

## DAFTAR PUSTAKA

- A Septiani and N Y Rustaman. (2017). Implementation of Performance Assessment in STEM (Science, Technology, Engineering, Mathematics) Education to Detect Science Process Skill. *IOP Conf. Series: Journal of Physics: Conf. Series* 812 (2017) 012052 hal 1-6 doi:10.1088/1742-6596/812/1/012052
- Abdullah, Sani Ridwan. (2014). *Pembelajaran saintifik untuk kurikulum 2013*. Jakarta: Bumi Aksara.
- Abdussakir. (2002). *Pembelajaran Geometri Berdasar Teori van Hiele Berbantuan Komputer*. Jurnal Matematika atau Pembelajarannya. Tahun VIII, Edisi Khusus: 344-348.
- Abdussakir, (2003). Pembelajaran Geometri Berdasarkan Teori van Hiele Berbantuan Komputer. *Jurnal Matematika atau Pembelajarannya*. Tahun VIII. Edisi Khusus
- Archbald, D.A., & Newmann, F.M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary school*. Washington, DC: Office of Educational Research and Improvement.
- Adams, R. J., Wilson, M. R., and Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23.
- Adams, R. J., Wilson, M., and Wu, M. (1997). Multilevel item response modeling: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22, 47-76.
- Andrich, D. (1998). Thresholds, steps and rating scale conceptualization. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 12, 648.
- Aday, LA., Begley, CE., Lairson, DR., Slater, CH., Richard, AJ., Montoya, ID. (1993). A Framework for Assessing the Effectiveness, Efficiency, and Equity of Behavioral Healthcare. *The American Journal Of Managed Care*, 5, 25- 44.
- Adiguzel, T. (2011). Use of Audio Modification in Science Vocabulary Assessment. *Eurasia Journal of Mathematics, Science, Technology Education*, 7(4): 215-225.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association (AERA).

Agustan. (2012). *Profil Berpikir Geometris Siswa SMP Level Deduksi Informal dalam Memahami Hubungan Antar bangun Segi empat Berdasarkan Gaya Belajar*. Tesis tidak diterbitkan. Surabaya: Program Pasca Sarjana Universitas Negeri Surabaya

Airasian, P. W., and Russel, M. K. (2008). *Classroom Assessment: Concepts and Applications* (6rd ed). New York: Mc. Graw Hill.

Akbar, S. (2013). *Instrumen Perangkat Pembelajaran*. Bandung: Remaja Rosdakarya.

Akiyama, T. (2003). Assessing speaking: Issues in school-based assessment and the introductionof speaking tests into the Japanese senior high school entrance examination. *JALT Journal*, 25, 117–141.

Allen & Yen,. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Publishing Company.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery ofassessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment*, 1(5). <http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml>

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing* . Washington, DC: AERA.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. *Psychological Bulletin*, 51(2, suppl.).

American Psychological Association [APA], American Educational Research Association [AERA], & National Council on Measurement in Education [NCME]. (1974). *Standards for educational and psychological tests and manuals*. Washington, DC: AERA.

Amir, Almira. (2014). Pembelajaran Matematika SMP Dengan Menggunakan Media Manipulatif. *Jurnal Forum Pedagogik*. 6 (1):78-79.

Ana, N., Sugiman. (2017). Efektivitas Pendekatan Matematika Realistik dalam Pembelajaran Matematika ditinjau dari Kemampuan Penalaran Siswa Kelas VII SMP Muhammadiyah 1 Sleman. Nurlatifah, Ana & Sugiman. 2017. Efektivitas Pendekatan Matematika Realistik Dalam Pembelajaran Matematika Ditinjau dari *Jurnal Pendidikan Matematika*, 6(5).

Anderson, L.W. (Ed.), Krathwohl, D.R. (Ed.), Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., & Wittrock, M.C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives (Complete edition)*. New York: Longman.

Andono, dkk. (2003). *Standar kompetensi bidang keahlian busana "Custom-made"* Jakarta: PPPG Kejuruan

Andrade, H.G. (2001). The effects of instructional rubrics on learning to write. *Current Issues in Education* 4, no. 4. <http://cie.ed.asu.edu/volume4/number4> (accessed January 4, 2019).

Andrade, H., and Y. Du. (2005). Student Perspectives on Rubric-referenced Assessment. *Practical Assessment, Research & Evaluation* 10 (3): 1–11

Andrade, Heidi, Colleen Buff, Joe Terry, Marilyn Erano, and Shaun Paolino. (2009). Assessment-driven Improvements in Middle School Students Writing. *Middle School Journal* 40 (4): 4–12.

Andrade, Heidi L., Xiaolei Wang, Ying Du, and Robin L. Akawi. (2009). “Rubric-referenced Self-assessment and Self-Efficacy for Writing.” *The Journal of Educational Research* 102 (4): 287–302.

Andrade, Heidi L., Du Ying, and Kristina Mycek. (2010). “Rubric-referenced Self-assessment and Middle School Students’ Writing.” *Assessment in Education: Principles, Policy & Practice* 17 (2): 199–214.

Andrade, H. (2010). *Students as the definitive source of formative assessment: Academic self-assessment and the self-regulation of learning*. In H. J. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 90–105). New York: Routledge.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.

- Andrich, D. (1998). Thresholds, steps and rating scale conceptualization. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 12, 648.
- Andrich, D., Sheridan, B. E., & Luo, G. (2004). *RUMM2020: Rasch unidimensional measurement models [Computer software]*. Perth, Western Australia: RUMM Laboratory.
- Angoff, W. H. (1993). *Perspectives on differential item functioning methodology*. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillside, NJ: Lawrence Erlbaum.
- Archbald, D., and F. M. Newmann. (1988). *Beyond Standardized Tests: Assessing Authentic Achievement in the Secondary School*. Reston, VA: National Association of Secondary Principals.
- Ardli, I., Gafar, A., & Mudjalipah, S. (2012). Perangkat Asesmen unjuk kerja Untuk Pembelajaran Teknik Pemeliharaan Ikan. *Jurnal INVOTEC*, 8(2): 147-166.
- Arifin, Z. (2011). *Evaluasi Pembelajaran*. PT Remaja Rosdakarya. Bandung.
- Arikunto, S. (2013). *Dasar-Dasar Evaluasi Pendidikan Edisi II*. Bumi Aksara. Jakarta.
- Arter, J., and J. McTighe. 2001. *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Thousand Oaks, CA: Corwin/Sage.
- Arter, J., and J. Chappuis. (2007). *Creating and recognizing quality rubrics*. Upper Saddle River, NJ: Pearson/Merrill Prentice Hall
- Ash, Sarah L, Patti H. Clayton, and Maxine P. Atkinson. (2005). Integrating Reflection and Assessment to Capture and Improve Student Learning. *Michigan Journal of Community Service Learning* 11 (2): 49–60.
- Aso, Y. (2000). A comparison of holistic and analytic scorings for oral interview tests. *ARELE*, 11,131–139.
- Astawa, I. N., Mantra, I. B. N., & Widiastuti, I. A. M. S. (2017). Developing Communicative English Language Tests for Tourism Vocational High School Students. *International Journal of Social Sciences and Humanities (IJSSH)*, 1(2), 58-64.

- Avanzino, Susan. (2010). Starting from Scratch and Getting Somewhere: Assessment of Oral Communication Proficiency in General Education across Lower and Upper Division Courses. *Communication Teacher* 24 (2): 91–110. doi:10.1080/17404621003680898.
- Azwar, S. (2012). *Reliabilitas dan Validitas*. Yogyakarta: Pustaka Pelajar
- Azwar, S. (2014). *Reliabilitas dan Validitas (Edisi IV)*. Yogyakarta: Pustaka Belajar.
- Bambang Subali. (2010). *Penilaian Evaluasi dan Remediasi Pembelajaran*. UNY. Yogyakarta.
- Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, 17, 1–42.
- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.
- Baghaei, P., & Amrahi, N. (2011). Validation of a Multiple Choice English Vocabulary Test with the Rasch Model. *Journal of Language Teaching and Research*, 2(5), 1052-1060.
- Baker, E. L., Abedi, J., Linn, R. L., and Niemi, D. (1996). Dimensionality and generalizability of domain-independent performance assessments. *Journal of Educational Research*, 89, 197-205.
- Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment & Evaluation in Higher Education*, 27(5), 427–441 <http://dx.doi.org/10.1080/0260293022000009302>.
- Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, 2, 49–58.
- Bassok, M. (2001). *Semantic alignments in mathematical word problems*. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 401–433). Cambridge, MA: MIT Press.

Basuki, I. & Hariyanto. (2016). *Asesmen Pembelajaran*. Bandung: Remaja RoSMPakarya.

Bathesta, Y, LD Wahyuni. (2011). *Rubrik: Asesmen Alternatif Untuk Menilai Peserta Didik Secara Realtime Dan Komprehensif*. Makalah disajikan dalam Konferensi Himpunan Evaluasi Pendidikan Indonesia (HEPI), Hotel Nusantara, Bandar lampung

Battista MT, Wheatley GW & Talsma G. (1989). Spatial visualization, formal reasoning, and geometric problem-solving strategies of preservice elementary teachers. *Focus on Learning Problems in Mathematics*, 11(4):17-30.

Beale, E. M. L., and Little, R. J. A. (1975). Missing data in multivariate analysis. *Journal of the Royal Statistical Society (B)*, 129-145.

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). *Human scoring*. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.),

Bergee, M. J. (2003). Faculty inter judge reliability of music performance evaluation. *Journal of Research in Music Education*, 51 (2), 137–150.

Bergen, D. (1993-1994). Authentic performance assessments. *Childhood Education*, 70(2), 99-102.

Bernardin, H. J., and Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology*, 65, 60-66

Berk, Ronald, A.(1986). *Performance Assessment*. London: The John Hopkins Press Ltd.

Berk, R.A. (1986). Preface. In R.A. Berk (Ed.), *Performance assessment: Methods & applications* (pp. ix-xiv). Baltimore, Maryland, John Hopkins University Press.

Biggs, J., and C. Tang. (2007). *Teaching for Quality Learning at University*. Maiden head:Open University Press.

Blaz, Deborah (2008). *Differentiated assessment for middle and high school classrooms*. Larchmont, New York: Eye on Education.

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

- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York, NY: Routledge.
- Boone, W., & Scantlebury, K. (2006). The role of Rasch analysis in science education utilizing multiple choice tests. *Science Educationm*, 90, 253-269.
- Kelly, T. L. (1927). *Interpretation of educational measurements*. New York: Macmillan.
- Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance*, 20, 238-252
- Boud, D., 1990. Assessment and the promotion of academic values. *Studies in Higher Education* 15 (1), 101e111
- Boud, D. (1995). "Assessment and Learning: Contradictory or Complementary?" In *Assessment for Learning in Higher Education*, edited by P. T. Knight, 35-48. London: Kogan Page.
- Boud, D., Cohen, R., & Sampson, J. (1999). Peer learning and assessment. *Assessment & Evaluation in Higher Education*, 24(4), 413–426  
<http://dx.doi.org/10.1080/0260293990240405>.
- Boud, D., and R. Soler. (2015). "Sustainable Assessment Revisited." *Assessment & Evaluationin Higher Education*. Advance online publication. doi:10.1080/02602938.2015.1018133.
- Box, C., Skoog, G., Dabbs, J.M, (2015). A Case Study of Teacher Personal Practice Assessment Theories and Complexities of Implementing Formative Assessment
- Brakel, T. D. (2006). Inter-judge reliability of the Indiana State School Music Association high school instrumental festival. *Journal of Band Research*, 42(1), 59–69.
- Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, 13, 1-18.
- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.
- Brennan, R; L. (1983).*Elements of Generoalizability Theory* (4<sup>th</sup>ed). Iow a City: ACT Publications.

- Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational measurement: Issues and practice*, 16 (4), 14–20.
- Brennan, R. L. (2000). (Mis) conceptions about generalizability theory. *Educational Measurement: Issues and Practice*, 19, 5-10
- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.
- Brennan, R. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24 , 1–21.
- Brian C. W., Stefanie A. W., &George E, Jr. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*. Vol. 19(2) 147–170. DOI: 10.1177/1029864915589014
- Brookhart, Susan M. (2013a). “The Public Understanding of Assessment in Educational Reform in the United States.” *Oxford Review of Education* 39: 52–71.
- Brown, S., and P. Knight. (1994). “Assessing Learners in Higher Education.” In Teaching and Learning in Higher Education, edited by J. Stephenson. London: Kogan Page.
- Brown, G., J. Bull, and M. Pendlebury. 1997. *Assessing Student Learning in Higher Education*. London: Routledge.
- Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology*, 20, 493–523.
- 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(2), 105–121 <http://dx.doi.org/10.1016/j.asw.2004.07.001>.
- Brown, G., and M. Craig. (2004). “Assessment of Authentic Learning.” Accessed September 13, 2018. <http://www.coe.missouri.edu/~vlib/glennglenn.michelle's.stuff/GLEN3MIC>
- Box, C., Skoog, G., Dabbs, J.M, 2015. A Case Study of Teacher Personal Practice Assessment Theories and Complexities of Implementing Formative Assessment. *American Educational Research Journal*

- Brown, A. (2003). *Interviewer variability in oral proficiency interviews*. Frankfurt, Germany: Lang.
- Bruce B. Frey, Vicki L. Schmitt, Justin P. Allen. 2012. Defining Authentic Classroom Assessment. *Practical Assessment, Research & Evaluation*, Vol 17, No 2
- B.Uno, Hamzah. (2008). *Model Pembelajaran*, Jakarta; PT. Bumi Aksara
- Buhagiar, M. A. (2007). Classroom assessment within the alternative assessment paradigm: revisiting the territory. *The Curriculum Journal*, 18(1), 39-56. DOI:10.1080/09585170701292174
- Burnsed, V., Hinkle, D., & King, S. (1985). Performance evaluation reliability at selected concert festivals. *Journal of Band Research*, 21 (1), 22–29.
- Burger, W.F. & Shaughnessy, J.M. (1986). Characterizing the Van Hiele Levels of Development in Geometry. *Journal for Research in Mathematics Education*, 31-47.
- Cason, G. J., and Cason, C. L. (1984). A deterministic theory of clinical performance rating. *Evaluation and the Health Professions*, 7, 221-247.
- Catalano, M . G., Perucchini, P., & Vecchio, G.M. (2014). The Quality Of Teachers' Educational Practices:.WCLTA 2013. *Procedia - Social and Behavioral Sciences* 141 (2014) 459–464.doi:10.1016/j.sbspro.2014.05.080
- Charoenchai, C., Phuseeorn, S. and Phengsawat,W. 2015. Teachers development modelto authentic assessment by empowerment evaluation approach. *Educational Research and Reviews*, 10(17), 2524-2530.
- Chasteen, Stephanie V., E. Rachel Pepper, Marcos D. Caballero, Steven J. Pollock, and Katherine K. Perkins. (2012). "Colorado Upper-division Electrostatics Diagnostic: A Conceptual Assessment for the Junior Level." *Physical Review Special Topics Physics Education Research* 8 (2): 020108. doi:10.1103/PhysRevSTPER.8.020108.
- Cheung, A.C.K. & Wong, P.M. 2012. Factor saffecting the implementation of curriculum reform in Hong Kong: key findings from a large-scale survey study. *International Journal of Educational Management*, 36 (1), 39-54

Christie, C. A. (2003) 'What Guides Evaluation? A Study of How Evaluation Practice Maps onto Evaluation Theory', *New Directions for Evaluation* 97: 7–35

Cho, Kwangsu, Christian D. Schunn, and Roy W. Wilson. (2006). Validity and Reliability of Scaffolded Peer Assessment of Writing from Instructor and Student Perspectives. *Journal of Educational Psychology* 98 (4): 891–901. doi:10.1037/0022-0663.98.4.891.

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

Clayton MJ. (1997). Delphi: a technique to harness expert opinion for critical decision making tasks in education. *Educ Psychol.*;17:373–386.

Collette, A.T. & Chiappetta, E. L. (1994). *Science Instruction in the Middle and Secondary Schools (3rd edition.)* New York: Merrill.

Conrad, D. (2003). Judging the judges: Improving rater reliability at music contests. *NFHS Music Association Journal*, 20 (2), 27–31.

Crisp, G. 2009. "Towards Authentic E-Assessment Tasks." In Ed Media World Conferenceon Educational Multimedia, Hypermedia and Telecommunications 2009, edited by G. Siemens and C. Fulford, 1585–1590. Chesapeake, VA:AACE.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart and Winston.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 16 (2), 137–163.

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

- Crowe, Edward. (2010). *Measuring What Matters A Stronger Accountability Model for Teacher Education*. The Center for American Progress. [www.americanprogress.org](http://www.americanprogress.org)
- Custer, R. L. At al. (2000). *Using Authentic Assessment in Vocational Education*. Clearinghouse on Adults, Career, and Vocational Education.The Ohaio State University
- D Fauziah, Mardiyana, D R S Saputro. (2018). Mathematics authentic assessment on statistics learning: the case for student mini projects. International Conference on Mathematics, Science and Education 2017 (ICMSE2017). *IOP Conf. Series: Journal of Physics: Conf. Series* 983 (2018) 012123. doi :10.1088/1742-6596/983/1/012123
- Darling-Hammond,L J. Snyder. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education* 16, 523- 545.
- Darling-Hammond, L. (2000). *Educating teachers*. San Francisco:Jossey Bass, forthcoming.
- Daryanto, (2010). *Media Pembelajaran Peranannya Sangat Penting Dalam Mencapai Tujuan Pembelajaran*, Yogyakarta: Gava Media.
- DeGruijter, D. N. M. (1984). Two simple models for rater effects. *Applied Psychological Measurement*, 8, 213-218.
- Djaali & Muljono, Pudji. (2008). Pengukuran dalam Bidang Pendidikan. Jakarta: Grasindo
- Depdiknas. (2006). *Peraturan Menteri No 23Tahun 2006 Tentang Standar Kompetensi Kelulusan*. Jakarta: Depdiknas
- Dikli, Semire. (2003) Assessment at a distance: Traditional vs. Alternative Assessments. *The Turkish Online Journal of Educational Technology- TOJET* July 2003 ISSN: 1303-6521 volume 2 Issue 3 Article 2.
- Diller, K. and Phelps, S. (2008). Learning outcomes, portfolios, and rubrics, oh my! Authentic assessment of an information literacy program. *Portal: Libraries and the Academy* 8(1), pp. 75-89. Available at: <http://dx.doi.org/10.1353/pla.2008.0000>.
- Diaz-Rico, Lynne T. (2008). *Strategies for Teaching English Learners*. Pearson Education, Inc

Du, Y., Wright, B. D., & Brown, W. L. (1996). *Differential facet functioning detection in direct writing assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Dungus, F. (2013). The Effect of Implementation of Performance Assessment, Portfolio Assessment and Written Assessments Toward the Improving of Basic Physics II Learning Achievement. *Journal of Education and Practice* Vol.4, No.14, 2013.

Duval R. (1998). *Geometry from a cognitive point of view*. In C Mammana & V Villani (eds). *Perspectives on the Teaching of Geometry for the 21st Century: An ICMI Study*. Dordrecht: Kluwer

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

Ebel, R.L. and Frisbie, D.A. (1991) *Essentials of Educational Measurement*. 5th Edition, Prentice-Hall, Englewood Cliffs.

Eckes, T. (2004). Beurteilerübereinstimmung und Beurteilerstrenge: Eine Multifacetten-Rasch-Analyse von Leistungsbeurteilungen im „Test Deutsch als Fremdsprache“ (TestDaF) [Rater agreement and rater severity: A many-facet Rasch analysis of performance assessments in the Test of German as a Foreign Language (TestDaF)]. *Diagnostica*, 50, 65–77.

Eckes, T. (2005a). Evaluation von Beurteilungen: Psychometrische Qualitätssicherung mit dem Multifacetten-Rasch-Modell [Evaluation of ratings: Psychometric quality assurance via many-facet Rasch measurement]. *Zeitschrift für Psychologie*, 213, 77–96.

Eckes, T. (2005b). Examining rater effects in Test DaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly* 2, 197–221.

Eckes, T. (2008a). *Assuring the quality of TestDaF examinations: A psychometric modeling approach*. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity – Proceedings of the ALTE Berlin Conference May 2005* (pp. 157–178). Cambridge, UK: Cambridge University Press.

Eckes, T. (2008b). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25, 155–185.

- Eckes, T. (2009). On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown & K. Hill (Eds.), Tasks and criteria in performance assessment. *Proceedings of the 28th Language Testing Research Colloquium* (pp. 43–73). Frankfurt, Germany: Lang.
- Elbow, P., & Yancey, K. B. (1994). On the nature of holistic scoring: An inquiry composed on e-mail. *Assessing Writing*, 1, 91–107.
- Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, 2, 175–196.
- Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, 24, 37–64.
- Ellis, Arthur K. (2001). *Teaching, learning, & assessment together. The reflective classroom*. Larchmont, New York: Eye On Education.
- Elshout-Mohr, M. et al., (2002). Student assessment within the context of constructivist educational settings. *Studies in Educational Evaluation* 28(4), pp. 369-390. Available at: [http://dx.doi.org/10.1016/S0191-491X\(02\)00044-5](http://dx.doi.org/10.1016/S0191-491X(02)00044-5)
- Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment* 8 (4),341–349.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists* . Mahwah, NJ: Erlbaum.
- Engelhard, G. (2013). Invariant measurement: Using Rasch models in the social, behavioral, and health sciences . New York: Routledge.
- Engelbrecht, P., Green, L., Naicker, S. & Engelbrecht, L. (1999). *Inclusive Education in Action in South Africa*. Pretoria: J.L. van Schaik.
- Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, 5(3), 171–191.
- Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31, 93-112.

- Engelhard, G., Jr. (1996a). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, 33(1), 56–70.
- Engelhard, G., Jr. (1996b). Examining rater errors in the assessment of written composition with a manyfaceted Rasch model. *Journal of Educational Measurement*, 33(2), 115–116.
- Engelhard, G., Jr. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, 1(1), 19–33.
- Engelhard, G., Jr. (2002). *Monitoring raters in performance assessments*. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for all students: Development, implementation, and analysis* (pp. 261–287). Mahwah, NJ: Erlbaum.
- Engelhard, G., Jr. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement*, 69(4), 585–602.
- Engelhard, G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge.
- Engelhard, G., Jr., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the advanced placement English literature and composition program with a Many-faceted Rasch Model* (ETS Research Report). Retrieved from [http://www.ets.org/research/policy\\_research\\_reports/publications/report/2003/ihav](http://www.ets.org/research/policy_research_reports/publications/report/2003/ihav)
- Engelhard, G., Jr., & Behizadeh, N. (2012). Epistemic iterations and consensus definitions of validity. *Measurement: Interdisciplinary Research and Perspectives*, 10(1), 55–58.
- Engelhard, G., Jr., & Wind, S. A. (2013). *Rating quality studies using Rasch measurement theory* (Research Report 2013-3). New York: The College Board. Retrieved from <https://research.collegeboard.org/sites/default/files/publications/2013/8/reseachreport-2013-3-rating-quality-studies-using-raschmeasurement-theory.pdf>
- Ercikan, K., Arim, R., Law, D., Domene, J., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning id entified by expert reviews. *Educational Measurement: Issues and Practice*, 29 , 24–35.

- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87 (1), 215–251.
- Fajar, A. (2009). *Portofolio dalam Pembelajaran*. PT Remaja RoSMPakarya Offset. Bandung.
- Fauzi, I. (2015). *Studi Deskripsi Implementasi Kurikulum 2013 pada Pembelajaran Fisika di Wilayah SMA Negeri Kabupaten Bantul*. Skripsi tidak diterbitkan. Yogyakarta: UIN Sunan Kalijaga.
- Fatimah. (2009). *Fun Math: Matematika Asyik dengan Metode Pemodelan*. Bandung: DAR! Mizan.
- Fiske, H. E. (1983). *The effect of a training procedure in music performance evaluation on judge reliability*. Toronto, Canada: Ontario Ministry of Education.
- Foley, R.C. dkk. (1972). *Dairy Cattle Principles, Practices, Problems, Profits*. LEA & Febiger. Philadelphia
- Fourie, I. and Niekerk, D. van. (1999). Using portfolio assessment in a module in research information skills. *Education for Information* 17(4), pp. 333-352.
- Fourie, I. and Niekerk, D. van. (2001). Follow-up on the use of portfolio assessment for a module in research information skills: an analysis of its value. *Education for Information* 19(2), pp. 107-126.
- Frederiksen, N. (1984). The real test bias: influences of testing on teaching and learning. *American Psychologist* 39(3), pp. 193-202. Available at: <http://dx.doi.org/10.1037/0003-066X.39.3.193>.
- Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18 (9), 27-32.
- Freedman, S. W., & Calfee, R. C. (1983). *Holistic assessment of writing: Experimental design and cognitive theory*. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75–98). New York: Longman
- Frey, B.B. & Schmitt, V.L. (2006, April). *Coming to terms with classroom assessment: Are performance-based assessments authentic or visa-versa?* Presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Frey, B.B. & Schmitt, V.L. (2007). Coming to terms with classroom assessment. *Journal of Advanced Academics*, 18, 3,402-423

Fluckiger, J. (2010). "Single Point Rubric: A Tool for Responsible Student Self-assessment".*Delta Kappa Gamma Bulletin* 76 (4): 18–25.

Fukazawa, M. (2010). Validity of peer assessment of speech performance. *ARELE*, 21, 181–190.

Gentner, D., Holyoak, K. J.,&Kokinov, B. (Eds.). (2001). *The analogical mind: Perspectives from cognitive science*. Cambridge, MA: MIT Press.

Gielen, S., Dochy, F., Dierick, S., (2003). *Evaluating the consequential validity of new modes of assessment: the influence of assessment on learning, including the pre-, post-, and the true assessment effects*. In: Segers,

Gipps, C. (1995) Reliability validity and manage ability in large scale performance assessment, *paper presented at AERA Conference*, Atlanta, 1993, in: H. TORRANCE (Ed.), Evaluating Authentic Assessment (Milton Keynes, Open University Press). 105-123

Glencoe, McGraw-Hill. (2006). *Performance Assessment In The Science Classroom*. Orion Place: USA

Goswami, U. (1992). *Analogical reasoning in children*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Goos M, Stillman G & Vale C. (2007). *Teaching Secondary School Mathematics: Research and Practice for the 21st Century*. Singapore: CMO Image Printing.

Gratch Lindauer, B. (2003). *Selecting and developing assessment tools*. In: Avery, E. (Ed.) *Assessing student learning outcomes for information literacy instruction in academic institutions*. Chicago: American Library Association, pp. 22-39.

Gronlund, G. (2003). *Focused early learning: A planning framework for teaching young children*. St. Paul, MN: RedleafPress.

Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S. A., Wooten, A., & Graham, M. (2012). Show ups versus lineups: An evaluation using ROC analysis. *Journal of Applied Research in Memory and Cognition*, 1, 221-228.

- Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw Hill.
- Gulikers, J., Bastiaens, T., & Kirschner, P. (2004). A five dimensional framework For authentic assessment. *Educational Technology Research & Development*, 52(3), 67-86.
- Gulikers, J., Bastiaens, T., & Martens, R. (2005). The surplus value of an authentic learning environment. *Computers in Human Behavior*, 21, 509-521.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley
- Hanna, Gerald S. and Dettmer, Peggy A. (2004). *Assessment for Effective Teaching*. Boston: Pearson Education Inc.
- Hagan, S. O., Pill, J., & Zhang, Y. (2016). Extending the scope of speaking assessment criteria in aspecific-purpose language test: Operationalizing a health professional perspective. *Language Testing*, 33, 195–216. doi:10.1177/0265532215607920
- Habron, Geoffrey; Goralnik, Lissy; Thorp, Laurie. (2012). Embracing the Learning Paradigm to Foster Systems Thinking. *International Journal of Sustainability in Higher Education*, v13 n4 p378-393 2012
- Hafner, O. C., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education*, 25 (12), 1509–1528 <http://dx.doi.org/10.1080/0950069022000038268>
- Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing*, 12, 1–9
- Hancock, D. R. (2004) ‘Cooperative Learning and Peer Orientation Effects on Motivation and Achievement’, *The Journal of Educational Research* 97(3): 159–66.
- Hash, P. M. (2012). An analysis of the ratings and interrater reliability of high school band contests. *Journal of Research in Music Education*, 60(1), 81–100.
- Havyer, R. D., Wingo, M. T., Comfere, N., Nelson, D. R., Halvorsen, A. J., McDonald, F. S., & Reed, D. A. (2014). Teamwork assessment in internal medicine: A systematic review of validity evidence and outcomes. *Journal of General Internal Medicine*, 29, 894-910.

- Hedge, J. W., and Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *Journal of Applied Psychology*, 73, 68-73.
- Hendrickson, A., & Yin, P. (2010). *Generalizability theory*. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 115–122). New York: Routledge.
- Henning, G. (1997). Accounting for nonsystematic error in performance ratings. *Language Testing*, 13(1), 53–63.
- Herrington, J., T. C. Reeves, and R. Oliver. (2010). *A Guide to Authentic E-Learning*. London: Routledge.
- Heruman. (2013). *Model Pembelajaran Matematika* (Boyke. R, Ed.) (Edisi ke-5, Cetakan ke-1). Bandung: PT Remaja RoSMPakarya Offset
- Hidriyah, N., & Wasis. (2014). Penerapan Self Assessment untuk Feedback pada Asesmen unjuk kerja Siswa dalam Kegiatan Praktikum Materi Fluida Statis Kelas XI SMA Negeri 1 Babat Lamongan. *Jurnal Inovasi Pendidikan Fisika (JIPF)*, 03, 60-66
- Hibbard, M. (1995). *Performance Assessment in the Science Classroom*. New York: The McGraw-Hill Companies.
- Himberg Cathrine & Hutchinson, Gayle E.&Roussell John M. 2003. *Secondary Physical Education. Preparing Adolescents to be Active for Life*. United State: Human Kinetics.
- Hill, C. E., O'Grady, K. E., and Price, P. (1988). A method for investigating sources of rater bias. *Journal of Counseling Psychology*, 35, 346-350
- Hoffer, A. (1981). Geometry is more than proof. *Mathematics Teacher*, 74, 11 -18.
- Hogan, Thomas P. (2007). *Educational Assessment A Practical Introduction*. John Wiley & Sons: USA
- Hopple, M. S. Christine J. 2005. *Elementary Physical Education Teaching & Assessment. A Practical Guide*. USA: Human Kinetics.
- Hoskens, M., and Wilson, M. (2001). Real-time feedback on rater drift in constructed response items: An example from the Golden State Examination. *Journal of Educational Measurement*, 38, 121-146.

- Houston, W. M., Raymond, M. R., and Svec, J. C. (1991). Adjustments for rater effects in performance assessment. *Applied Psychological Measurement*, 15, 409-421.
- Howell, Rebecca J. (2011). "Exploring the Impact of Grading Rubrics on Academic Performance: Findings from a Quasi-experimental, Pre-post Evaluation." *Journal on Excellence in College Teaching* 22 (2): 31–49.
- Hoyt, C. J. (1941). *Test reliability estimated by analysis of variance*. *Psychometrika*, 6, 153–160.
- Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, 4, 403–424.
- Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it?. *Psychological Methods*, 5, 64–86.
- Husamah dan Y. Setyaningrum. (2013). *Desain Pembelajaran Berbasis Pencapaian Kompetensi*. Prestasi Pustakaraya. Jakarta.
- Hutabarat, O. R. (2004). *Model-model Penilaian Berbasis Kompetensi PAK*. Bandung: Bina Media Informasi.
- Ibrahim M & M. Nur dalam Rusman. (2012). *Pembelajaran Berdasar Masalah*. hlm.243. Surabaya: UNESA University Press.
- Izza, L.N. (2014). *Analisis Instrumen Performance Assessment dengan Metode Generalizability Coefficient Pada Ketermpilan Dasar Laboratorium*. *Jurnal Chemistry in Education*, 3(1): 30-36
- Jaedun, Amat. (2010). *Metode Penelitian Evaluasi Program. Evaluasi Kebijakan dan Evaluasi Program Pendidikan*. Lembaga Penelitian Universitas Negeri Yogyakarta. Yogyakarta
- Jensen, Eric (2005). *Teaching with the brain in mind*, 2nd Edition. Alexandria, Virginia: Association for Supervision and Curriculum Development.
- Jihad, A. dan A. Haris. (2012). *Evaluasi Pembelajaran*. Multi Pressindo. Yogyakarta
- Johnson, R. L., J. Penny, and B. Gordon. (2000). "The Relation between Score Resolution Methods and Interrater Reliability: An Empirical Study of an

- Analytic Scoring Rubric. "Applied Measurement in Education 13 (2): 121–138. doi:10.1207/S15324818AME1302\_1.
- Jonsson, A. and Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review* 2(2), pp. 130-144. Available at: <http://dx.doi.org/10.1016/j.edurev.2007.05.002>.
- Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York: Guilford.
- Jonsson, Anders. (2010). "The Use of Transparency in the 'Interactive Examination' for Student Teachers." *Assessment in Education: Principles, Policy & Practice* 17 (3): 183–197.
- Johnson, Elaine B. (2010). *CTL Contextual Teaching &Learning*. Bandung: Kaifa Learning
- Joreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2, pp. 1–56). San Francisco: W. H. Freeman.
- Karyana, I. 2013. Pengembangan Instrumen Penilaian Unjuk Kerja (Performance Assessment) Keterampilan Penerapan Metode Ilmiah dalam Penyusunan Skripsi Karya Seni Mahasiswa Program Studi Seni Rupa Murni Institut Seni Indonesia Denpasar, *MUDRA*. ISSN, Vol 28, Nomor 2, pp. 216-229.
- Khan, B. (2012). Relationship between assessment and students learning. *International Journal of Social Sciences and Education*, 2(1), 576-588.
- Kane, M. B., Khattri, N., Reeve, A. L., & Adamson, R. J. (1997). *Assessment of student performance*. Washington, DC: Studies of Education Reform, Office of Education Research and Improvement, U.S. Department of Education.
- Kane, M., Crooks, T. & Cohen, A. (1999) 'Validating Measures of Performance', *Educational Measurement: Issues and Practices* 18(2): 5–17.
- Keaveny, T. J., and McGann, A. F. (1975). A comparison of behavioral expectation scales and graphic rating scales. *Journal of Applied Psychology*, 60, 695-703.

- Kenny, D. A., and Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, 112, 165-172.
- Kennedy, M.L., Tipps, S., & Johnson A. (2008). *Eleventh edition guiding children's learning of mathematics*. Belmont:Thomson Wadsworth.
- Ketut, I.S. (2012). *Pengembangan Instrumen Penilaian Unjuk Kerja (Performance Assessment) Laboratorium pada Mata Pelajaran Fisika sesuai KTSP SMA Kelas X di Kabupaten Gianyar*. Skripsi tidak diterbitkan. Gianyar; UNDHINKSA.
- Kenderski, C. M. (1983). *Interaction process and learning among third-grade black and Mexican American students in cooperative small groups*. Unpublished doctoral dissertation, University of California, Los Angeles.
- Knight, L. 2006. Using rubrics to assess information literacy. *Reference Services Review* 34(1), pp. 43-55. Available at: <http://dx.doi.org/10.1108/00907320610640752>.
- Knoch, U., Read, J., & von Randow, J. (2007). *Re-training writing raters online: How does it compare with face-to-face training?*. *Assessing Writing*, 12, 26-43.
- Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, 26, 275-304.
- King, S. E., & Burnsed, V. (2007). A study of the reliability of adjudicator ratings at the 2005 Virginia band and orchestra directors association state marching band festivals. *Journal of Band Research*, 27-33.
- Kocakülah, Mustafa Sabri. (2010). "Development and Application of a Rubric for Evaluating Students' Performance on Newton's Laws of Motion." *Journal of Science Education and Technology* 19 (2): 146-164. doi:10.1007/s10956-009-9188-9.
- Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19(1), 3-31.
- Kosasih. (2014). *Strategi Belajar dan Pembelajaran Implementasi Kurikulum 2013*. Bandung: Yrama Widya.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Princeton, NJ: Princeton University Press.

Kusaeri. (2014). *Acuan & Teknik Penilaian Proses & Hasil Belajar dalam Kurikulum 2013*. Yogyakarta: Ar Ruzz Media.

Lajoie, S. (1991). A framework for authentic assessment in mathematics. *NCRMSE Research Review: The Teaching and Learning of Mathematics*, 1(1), 6-12.

Lance, C. E., LaPointe, J. A., and Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology*, 79, 332-340

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

Lane, S., & Stone, C. A. (2006). *Performance assessment*. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). Westport, CT: American Council on Education/Praeger.

Lane, Suzanne, and Sean T. Tierney. (2008). “*Performance Assessment*.” In *21st Century Education: A Reference Handbook* (Vol. 1), edited by Thomas L. Good, 461–470. Los Angeles, CA: SAGE.

Latham, G. P., Wexley, K. N., and Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology*, 60, 550-555.

Laudan, L. (1977). Progress and its problems: Toward a theory of scientific change . Berkeley: University of California Press

Lawrence M. Rudner, William D. Schafer. (2002). *What Teachers Need to Know about Assessment*. Washington, DC: National Education Association.

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>

Lee, Eunjung, and Sohyun Lee. (2009). “Effects of Instructional Rubrics on Class Engagement Behaviors and the Achievement of Lesson Objectives Typical Peers.” *Education and Training in Developmental Disabilities* 44 (3): 396–408.

- Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, 23(4), 6–15.
- Leo, I. D., Alivernini, F., & Lucidi, F. (2015). Psychometric properties and validity of an instrument measuring. 6th World conference on Psychology Counseling and Guidance, 14 - 16 May 2015. *Procedia Social and Behavioral Sciences* 205 (2015) 173 – 177. doi:doi:10.1016/j.sbspro.2015.09.053
- Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago: MESA Press.
- Linacre, J.M. and Wright, B.D. (1993): *A User's guide to FACETS: Rasch measurement computer program*. Version 3.2. Chicago, IL: MESA Press.
- Linacre, J.M. (1994): *Many-facet Rasch measurement*. Chicago, IL: MESA Press.
- Linacre, J. M. (1996). Generalizability and manyfacet Rasch measurement. In G. Engelhard, Jr., and M. Wilson (Eds.), *Objective measurement: Theory into practice*: Vol. 3 (pp. 8598). Norwood, NJ: Ablex.
- Linacre (1997): *MESA research note 2*. Unpublished paper presented at Midwest Objective Measurement Seminar, Chicago, IL, June, 1997.
- Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, 3, 103-122.
- Linacre, J.M. and Williams, J. (1999): *How much is enough? Rasch Measurement Transaction* 12, 653.
- Linacre, J. M. (2001a). *FACETS* [Computer program, version 3.36.2]. Chicago: MESA Press.
- Linacre, J. M. (2001b). *A user's guide to Facets: Rasch measurement computer program* [Computer program manual]. Chicago: MESA Press.
- Linacre, J. M. (2002). Number of person or item strata. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 16, 888.
- Linacre, J. M., & Wright, B. D. (2004). *Construction of measures from many-facet data*. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theories, models, and applications* (pp. 296–321). Maple Grove, MN: JAM Press.

Linacre J. M. (2012). *A User's Guide to Winsteps/Ministeps Rasch Model Computer Programs*.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment:Expectations and validation criteria. *Educational Researcher*, 20(8), 5-21 .

Linn, R. L., Burton, E., DeStefano, L., and Hanson, M. (1996). Generalizability of New Standards Project 1993 pilot study tasks in mathematics. *Applied Measurement in Education*, 9, 201-214.

Lin, C.P., et.al. (2011). The impact of using synchronous collaborative virtual tagram in childern's geometric. *TOJET: The Turkish Online Journal of Educational Technology*, 10(2). Diakses tanggal 11 Januari 2019 dari <http://eric.edu.gov>.

Linn, R. L. & Gronlund, N. E. (2005) *Measurement and Assessment in Teaching* (9th edn). Englewood Cliffs, NJ: Prentice Hall.

Linstone HA, Turoff M, eds. (1977). *The Delphi Method: Techniques and Applications*. London, UK: Addison-Wesley Publishing Company;

Little, R. J. A., and Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley.

Loeffler, Kelly A. (2005). "No More Friday Spelling Tests? An Alternative Spelling Assessment for Students with Learning Disabilities." *Teaching Exceptional Children* 37 (4): 24–27

Longford, N. T. (1994a). Reliability of essay rating and score adjustment. *Journal of Educational and Behavioral Statistics*, 19, 171-201.

Longford, N. T. (1994b). *A case for adjusting subjectively rated scores in the Advanced Placement tests (ETS Technical Report 945)*. Princeton, NJ: Educational Testing Service.

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

Longford, N. T. (1996). *Adjustment for reader rating behavior in the Test of Written English (TOEFL Research Report No. 55)*. Princeton, NJ: Educational Testing Service.

- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12, 54–71.
- Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters. *Language Testing*, 19(3), 246–276.
- Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt am Main, Germany: Peter Lang.
- Lund, J. 1997. "Authentic Assessment: Its Development and Applications." *Journal of Physical Education, Recreation, and Dance* 68 (7): 25–28.
- Lund, J.L.& D.Tannehill. 2005. *Standarts-Base Physical Education Curriculum Devolopment*. London: Jones and Bartlett Publishers.
- Lund, J.L & F.M. Kirk. 2010. *Performance Based Assessment for Middle and High School Physical Education*. United State:Human Kinetics.
- Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions* , 13 , 425–444.
- Lunz, M. E., Stahl, J. A., & Wright, B. D. (1996). The invariance of rater severity calibrations. In G. Engelhard, Jr. & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 99–112). Norwood, NJ: Ablex.
- Lunz, M., & Suanthong, S. (2011). Equating of multi-facet tests across administrations. *Journal of Applied Measurement* , 12 (2), 124–134.
- Lynch, B.K. and McNamara, T.F. 1998: Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing* 15, 158–80.
- M., Dochy, F., Cascallar, E. (Eds.), *Optimising New Modes of Assessment: In Search of Quality and Standards*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 37e54

- Macfarlane, B. (2011). The Morphing of Academic Practice: Unbundling and the Rise of the Para-academic. *Higher Education Quarterly* 65 (1): 59–73. doi:10.1111/j.1468-2273.2010.00467.x.
- MacMillan, P. D. (2000). Classical, generalizability, and multifaceted Rasch detection of interrater variability in large, sparse data sets. *Journal of Experimental Education*, 68, 167-190.
- Madeamin, Ishaq. ‘Objektivitas Dan Kepraktisan’, 2011. <http://shahibul1628.wordpress.com/2012/04/12/kepraktisan-dan-efek-potensial/>
- Mahmudah, S. (2000). *Penerapan Penilaian Kinerja Siswa (performance Assessment) pada Pembelajaran Sub Konsep Jaringan Hewan*. Bandung:UPI
- Majid, A. (2006). *Perencanaan Pembelajaran Mengembangkan Standar Kompetensi Guru*. Bandung: Remaja Rosdakarya.
- Manson, J. R., & Olsen, R. J. (2010). Diagnostics and rubrics for assessing learning across the computational science curriculum. *Journal of Computational Science*, 1(1), 55–61.
- Marcoulides, G. (1995). Generalizability theory and applications. Workshop presented at the 1995 *Language Testing Research Colloquium*, 24 March; Long Beach, CA.
- (1996). Estimating variance components in generalizability theory: the covariance structure analysis approach. *Structural Equation Modeling* 3, 290–99.
- Marcoulides, G. and Drezner, Z. (1996): *A method for analyzing performance assessment*. In Wilson, M., Draney, K. and Engelhard, G., editors, *Objective measurement: theory into practice*, Norwood, NJ: Ablex.
- Marcoulides, G. A., and Drezner, Z. (1997). A method for analyzing performance assessments. In M. Wilson, G. Engelhard, Jr., and K. Draney (Eds.), *Objective measurement: Theory into practice*: Vol. 4 (pp. 261-277). Greenwich, CT: Ablex.
- Marcoulides, G. A. (1999). Generalizability theory: Picking up where the Rasch IRT model leaves off? In S. E. Embretson, and S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 129-152). Mahwah, NJ: Lawrence Erlbaum

Mardapi, J. (2008). *Pengukuran Penilaian dan Evaluasi Pendidikan*. Yogyakarta :Nuha Medika

Mardapi, J. (2012). *Pengukuran Penilaian dan Evaluasi Pendidikan*. Yogyakarta :Nuha Medika

Marhaeni, A.A.I.N. (2008). *About Meaningfulness and Usefulness of Language Assessment. Presentedon the Regional Teflin Organization by the Faculty of Art*, Udayana University, July 1 5th-1 6th,201 2. (Retrieved from [http://www.undiksha.ac.id/e-learning/staff/images/img\\_info/4/1 8-282.pdf](http://www.undiksha.ac.id/e-learning/staff/images/img_info/4/1 8-282.pdf))

Marilyn H. Oermann, PhD, RN, ANEF, FAAN,. Kathleen B. Gaberson, PhD, RN, CNOR, CNE, ANEF. (2014). *Evaluation and Testingin Nursing Education*. Springer Publishing Company

Marsh, H. W. (1989). Confirmatory factor analysis of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, 13, 335-361.

Marsh, H. W., and Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement*, 15, 4770.

Martin, M. O. and D.L. Kelly, eds. (1996). *Third international mathematics and science study technical report*, volume I: design and development. Chestnut Hill, MA: Boston College. [Online]. <http://timss.bc.edu/timss1995i/TechVol1.html>. (Accessed 16 April 2019).

Martin, M.O., I.,V.S. Mullis and S.J. Chrostowski. (2004). *TIMSS 2003 technical report*. Chestnut Hill, MA:Boston College.

Masrukan. 2013. *Asesmen Otentik Pembelajaran Matematika*. Semarang: FMIPA Universitas Negeri Semarang

McNamara, T.F. (1996). *Measuring second language performance*. London: Longman.

McNamara, T.F. and Adams, R.J. (2000). *The implications of halo effects and item dependencies for objective measurement*. In Wilson, M. and Engelhard, G., editors, *Objective Measurement: theory into practice*. Vol. 5. Norwood, NJ: Ablex, 243–57.

McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.

McTighe, J. & Ferrata. 1998. *Assessing Learning in Classroom*. Website: <http://www.-msd.net/Assessment/authenticassessment.html>. Diunduh 23 April 2019.

Meenakshi, G. (2013). An assessment of final year project using fuzzy logic. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(9), 3392–3394.

Mellenbergh, G. J. (1977). The replicability of measures. *Psychological Bulletin*, 84 (2), 378. <http://doi.org/10.1037/0033-2909.84.2.378>

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

Messick, S. (1996) Validity of Psychological Assessment. Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist*, 5D, 741-749.

Mertler, C. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation* 7(25). [Online] Available at: <http://pareonline.net/getvn.asp?v=7&n=25> [Accessed: 1 Juni 2019].

Mertler, C. A. (2001). *Using Performance Assessment in Your Classroom*. Unpublished manuscript., Bowling Green State University.

Mert, S., & Karaca, D. (2010). The attitude of the prospective mathematics teachers toward instructional technologies and material development course. *TOJET: The Turkish Online Journal of Educational Technology* (9) 1. Diunduh tanggal 11 Januari 2019 dari [www.proquest.com](http://www.proquest.com).

Meyer, C. (1992). What's the difference between authentic and performance assessment? *Educational Leadership*, 49(8), 3940.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, 33(4), 379-416.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Johnson, L., & Almond, R. A. (2002). Making sense of data from complex assessments. *Applied Measurement in Education*, 15, 363–378.

- Mislevy, R. J., & Riconscente, M. M. (2006). *Evidence-centered assessment design: Layers, concepts, and terminology*. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, NJ: Erlbaum.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague/Mouton/Berlin: De Gruyter.
- Mosier, C. I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review*, 47, 355–366.
- Mosier, C. I. (1941). Psychophysics and mental test theory II: The constant process. *Psychological Review*, 48, 235–249.
- Moskal, Barbara M. (2000). Scoring Rubrics: What, When, How?. *Practical Assessment, Research, and Evaluation*. Tersedia: <http://ericace.net/pare/getun.asp?v=7&n=3>. [Rabu, 26 Desember 2019 pukul 21.02].
- Mott, Michael S., Debby A. Chessin, William J. Sumrall, Angela S. Rutherford, and VirginiaJ. Moore. (2011). “Assessing Student Scientific Expression Using Media: the Mediaenhanced Science Presentation Rubric (MESPR).” *Journal of STEM Education: Innovations and Research* 12 (1&2): 33–41.
- Mueller, J. (2016). The Authentic Asessment Toolbox. (Online), (<http://jolt.merlot.org/vol1no1/mueller.htm>), diakses 10 Desember 2018
- Mulyatiningsih, Endang. 2011. *Metode Penelitian Terapan Bidang Pendidikan*. Bandung: Alfabeta.
- Mulyatiningsih, Endang .2016. *Metode Penelitian Terapan Bidang Pendidikan*. Bandung: Alfabeta
- Mulyasa, E. (2013). *Pengembangan dan Implementasi Kurikulum 2013*. Cimahi: Rosda
- Muraki, E. (1999, April). *The introduction of essay questions to the GRE: Toward a synthesis of item response theory and generalizability theory*. Paper presented at the annual meeting of the American Educational Research Association, Montréal, Canada.
- Muraki, E., and Bock, R. D. (2003). *PARSCALE* [Computer program, version 4]. St. Paul MN: Assessment Systems Corporation.

Murphy, K. R., and Anhalt, R. L. (1992). Is halo error a property of the rater, ratees, or the specific behavior observed?. *Journal of Applied Psychology*, 77, 494-500.

Murphy, K. R., and DeShon, R. (2000a). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, 53, 873-900.

Murphy, K. R., and DeShon, R. (2000b). Progress in psychometrics: Can industrial and organizational psychology catch up?. *Personnel Psychology*, 53, 913-924.

Muslich, M. (2009). *KTSP (Kurikulum Berbasis Kompetensi)*. Jakarta: PT Bumi Aksara.

Mustapha, A., Samsudin, N. A., Arbaiy, N., Mohamed, R., & Rahmi, I. (2016). Generic assessment rubrics for computer programming courses. *Turkish Online Journal of Educational Technology*, 15(1), 53

Myford, C. M., & Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system* (No. ETS Center for Performance Assessment Report No. MS 94-05). Princeton, NJ: Educational Testing Service.

Myford, C. M., Marr, D. B., and Linacre, J. M. (1996). *Reader calibration and its potential role in equating for the TWE* (TOEFL Research Report No. 95-40). Princeton, NJ: Educational Testing Service.

Myford, C. M., & Wolfe, E. W. (2000). *Strengthening the ties that bind: Improving the linking network in sparsely connected rating designs* (TOEFL Technical Report, TR-15). Princeton, NJ: Educational Testing Service.

Myford, C. M., & Wolfe, E. W. (2002). When raters disagree, then what? Examining a third-rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings. *Journal of Applied Measurement*, 3, 300–324.

Myford, C. M., & Wolfe, E. W. (2003). *Detecting and measuring rater effects using many-facet Rasch measurement: Part I*. *Journal of Applied Measurement*, 4(4), 386–422.

- Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5(2), 189–227.
- Nana Syaodih Sukmadinata. (2005). *Metode Penelitian Pendidikan*. Bandung : Remaja Rosdakarya.
- National Council of Teachers of Mathematics. (1989). *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA Author.
- National Council on Education Standards and Testing (1992). *Raising standards for American Education*. Washington, DC: Author.
- NCTM. (2000). *Principles and standards for school mathematics*. Reston, VA: The National Council of Teachers of Mathematics, inc.
- Newmann, F. M. (1990). Higher order thinking in teaching social studies: A rationale for the assessment of classroom thoughtfulness. *Journal of Curriculum Studies*, 22(1), 41-56.
- Newmann, F., Brandt, R.. & Wiggins, G. (1998). An exchangeof views on semantics, psychometrics, and assessment reform: a close look at 'authentic' assessments. *Educational Researcher*, 27(6), 19-22.
- Newmann, Lori R., Beth A. Lown, Richard N. Jones, Anna Johansson, and Richard M.Schwartzstein. (2009). "Developing a Peer Assessment of Lecturing Instrument: Lessons Learned." *Journal of the Association of American Medical Colleges* 84 (8): 1104–1110.
- Nitko, A. J. (2001). *Educational Assessment of Students (3rd ed)*. Upper Saddle River, NJ: Merrill.
- Nitko, A. J. (2004). *Educational Assessment of Students (4th edn)*. Upper Saddle River, NJ: Merrill Prentice Hall.
- Nitko, A. and Brookhart, S. (2007). *Educational assessment of students*. 5th ed. Upper Saddle River, N.J.: Pearson Education.
- Nieven N and Folmer N (2013). *Formative Evaluation in Educational Design Research* (Netherlands Institute for Curriculum Development (SLO): Enschede) 152-169.

NCLRC, (2004). *Assessing Learning, Alternative Assessment. The National Capital Language Resource Center*. Washington, DC. Retrieved from <http://www.nclrc.org/essentials/assessing/alternative.htm>

Nordrum, L., K. Evans, and M. Gustafsson. (2013). "Comparing Student Learning Experiences of In-text Commentary and Rubric-articulated Feedback: Strategies for Formative Assessment." *Assessment & Evaluation in Higher Education* 38 (8): 919–940. doi:10.1080/02602938.2012.758229

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3 (1), 1–18.

Novick, L. R., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 398–415.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Nurgiyantoro, Burhan. (2008). "Penilaian Otentik". *Cakrawala Pendidikan Jurnal Ilmiah Pendidikan*. November. Th.XXVI, No. 3, Hal: 250-261. Doi: 10.21831/cp.v3i3.320

Nurkancana, Wayan. (1986). *Evaluasi Pendidikan*. Surabaya: Usaha Nasional

O'Grady, K. E., and Medoff, D. R. (1991). Rater reliability—a maximum-likelihood confirmatory factor-analytic approach. *Multivariate Behavioral Research*, 26, 363-387.

O'Neill, T. R., and Lunz, M. E. (2000). A method to study rater severity across several administrations. In M. Wilson, and G. Engelhard, Jr. (Eds.), *Objective measurement: Theory into practice*: Vol. 5 (pp. 135-146). Stamford, CT: Ablex.

Oakleaf, M. (2008). Dangers and opportunities: a conceptual map of information literacy assessment approaches. *Portal: Libraries and the Academy* 8(3), pp. 233-253. Available at: <http://dx.doi.org/10.1353/pla.0.0011>.

Oakleaf, M. (2009). Using rubrics to assess information literacy: an examination of methodology and interrater reliability. *Journal of the American Society for Information Science and Technology* 60(5), pp. 969-983. Available at: <http://dx.doi.org/10.1002/asi.21030>

- Oermann, M. H., & Gaberson, K. B. (2014). *Evaluation and testing in nursing education* (4th ed.). New York: Springer
- Offirston, Topic. (2014). *Aktivitas Pembelajaran Matematika Melalui Inkuiiri Berbantuan Software Cinderella*. Yogyakarta: Deepublish
- OKS. (2009). *Ortaogretim Kurumlari Ogrenci Secme ve Yerlestirme Sinavi Sayisal Bilgileri (The statistics for 2009 secondary education institutions student selection and placement examination)* [Online]. <http://oges.meb.gov.tr/oks/ista.html>. (Accessed 13 April 2019).
- Oktriawan, T. (2015). Pengembangan Instrumen Asesmen Kinerja Pada Praktikum Pengaruh Luas Permukaan Terhadap Laju Reaksi. *Jurnal Pendidikan dan Pembelajaran Kimia*, 4(2): 593-604.
- O'Malley, J.Michael & Pierce, Lorraine Valdez. (1996). *Authentic Assessment for English Language Learners*. USA: Addison-Wesley Publishing
- O'Neill, T. R., & Lunz, M. E. (2000). A method to study rater severity across several administrations. In M. Wilson & G. Engelhard, Jr. (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp.135–146). Stamford, CT: Ablex.
- Onwuegbuzie, A. J. (2000). Attitudes Toward Statistics Assessments. *Assessment & Evaluation in Higher Education* 25(4): 321–39.
- Orpwood, G. (2001). The Role of Assessment in Science Curriculum Reform. *Assessment in Education*, Vol. 8, No. 2, 2001
- O'Sullivan, B., & Rignall, M. (2007). *Assessing the value of bias analysis feedback to raters for the IELTS Writing Module*. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 446–478). Cambridge,UK: Cambridge University Press.
- Ozerem, A. (2012). "Misconceptions in Geometry and Suggested Solutions for Seventh Grade Students". *International Journal of New Trends in Arts, Sports and Science Education* . Vol. 1 No.4, pp. 23-35.
- Palm, T. (2008). Performance Assessment and Authentic Assessment: A Conceptual Analysis of The Literature. *Practical Assessment. Research & Evaluation*, 13 (4).(Online). (<http://pareonline.net/pdf/v13n4.pdf>), diakses 21 Desember 2018

Panadero, Ernesto, Jesús Alonso Tapia, and Juan Antonio Huertas. (2012). “Rubrics and Self-assessment Scripts Effects on Self-regulation, Learning and Self-efficacy in Secondary Education.” *Learning and Individual Differences* 22 (6): 806–813. doi:10.1016/j.lindif.2012.04.007.

Panadero, E., and A. Jonsson. (2013). “The Use of Scoring Rubrics for Formative Assessment Purposes Revisited: A Review.” *Educational Research Review* 9: 129–144.

Panadero, Ernesto, Jesús Alonso-Tapia, and Eloísa Reche. (2013). “Rubrics Vs. Self-assessment Scripts Effect on Self-regulation, Performance and Self-efficacy in Pre-service Teachers.” *Studies in Educational Evaluation* 39: 125–132.

Panadero, E., and M. Romero. (2014). “To Rubric or not to Rubric? The Effects of Self-assessment on Self-regulation, Performance and Self-efficacy.” *Assessment in Education: Principles, Policy & Practice* 21 (2): 133–148. doi:10.1080/0969594X.2013.877872

Pantiwati, Yuni. (2010). *Pengaruh Jenis Asesmen Biologi dalam Pembelajaran Kooperatif TPS (Think Pair Share) Terhadap Kemampuan Kognitif, Berpikir Kritis, Berpikir Kreatif, dan Kesadaran Metakognitif Siswa SMA di Kota Malang*. Desertasi tidak diterbitkan. Pascasarjana Universitas Negeri Malang.

Parke, C.S. (2001). An approach that examines sources of misfit to improve performance assessment items and rubrics. *Educational Assessment* 7, no. 3: 201–25

Parke,dkk. (2003). *Using Assessment to Improve Middle*. California

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

Patz, R. J., Wilson, M. J., and Hoskens, M. (1997). *Optimal rating procedures and methodology for NAEP open-ended items* (Working Paper No. 97-37). Washington, DC: U. S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics. Retrieved Sept. 5, 2001 from the World Wide Web: <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=9737>

Permendikbud RI No.65 Tahun 2013 tentang Standar Proses.

Permendikbud RI No.66 Tahun 2013 tentang Standar Penilaian

Permendikbud RI No. 103 Tahun 2014 tentang Pembelajaran pada Pendidikan Dasar dan Pendidikan Menengah.

Permendikbud RI No.104 Tahun 2014 tentang Standar Penilaian

Permendikbud RI No. 53 Tahun 2015 tentang Penilaian Hasil Belajar oleh Pendidikan Satuan Pendidikan pada Pendidikan Dasar dan Pendidikan Menengah.

Permen RI No.13 Tahun 2015 tentang Standar Nasional Pendidikan.

Popham, W. James. 1995. *Classroom Assessment: What Teacher Need to Know.*

Popham, W. James & Baker, Eva L.. 2008. *Teknik Mengajar Secara Sistematis.* Terjemahan Amirul Hadi, dkk. Jakarta: Rineka Cipta

Probosari, A. P. (2014). Analisis Deskriptif Penilaian Pembelajaran Keahlian Teknik Komputer Jaringan SMA di Kudus. *Seminar Nasional Evaluasi Pendidikan.* [online]. <http://conf.unnes.ac.id/index.php/snep/II/paper/view/191//84>. Diakses pukul 3.24 pm tanggal 4 Januari 2019.

Puckett, M.B., & Black, J.K. (2008). *Authentic assessment of the young child: celebrating development and learning* (2nd ed.). Des Moines, IA: Prentice-Hall Inc.

Pulakos, E. D., Schmitt, N., and Ostroff, C. (1986). A warning about the use of a standard deviation across dimensions within raters to measure halo. *Journal of Applied Psychology*, 71, 29-32.

Purwanto.(2014). *Evaluasi Hasil Belajar.* Yogyakarta: Pustaka Pelajar.

Puspawati. (2014). Pengaruh Pendekatan Kontekstual Berbantuan Asesmen Autentik Terhadap Hasil Belajar Matematika Ditinjau Dari Kemampuan Numerik Pada Siswa Kelas V SD Negeri 6 Gianyar Dan SD Negeri 7 Gianyar Di Gugus 1 Kecamatan Gianyar. *Volume 4 Tahun 2014 (Online).* <http://id.portalgaruda.org/index.php?ref=browse&mod=viewarticle&article=259268>. Diakses 12 Januari 2019

Prastowo, Andi. (2014). Pemenuhan Kebutuhan Psikologis Peserta Didik SD/MI Melalui Pembelajaran Tematik-Terpadu. *JPSD:Jurnal Pendidikan SMP.*

(Online) 1 (1):6 (<http://eprints.uny.ac.id>), diakses pada tanggal 20Februari 2019

- Puspitasari, N., Widiarti, N., & Haryani, S.(2014). *Pengembangan Rubrik Performance Assessment Pada Praktikum Hidrolisis Garam*. Skripsi. Semarang: Universitas Negeri Semarang.
- Race, P., S. Brown, and B. Smith. (2005). *500 Tips on Assessment*. 2nd ed. London:Routledge.
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research
- Rasch, G. (1961). *On general laws and meaning of measurement in psychology*. In J. Neyman (Ed.), Proceedings of the fourth Berkeley Symposium on mathematical statistics and probability (pp. 321–333). Berkeley: University of California Press.
- Rasch, G. (1977). *On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements*. Danish Yearbook of Philosophy 14, 58–94.
- Raymond, M. R. (1986). Missing data in evaluation research. *Evaluation and the Health Professions*, 9, 395-420.
- Raymond, M. R., and Houston, W. H. (1990). *Detecting and correcting for rater effects in performance assessment* (ACT Research Report Series 90-14). Iowa City, IA: The American College Testing Program.
- Raymond, M. R., Webb, L. C., & Houston, W. M. (1991). Correcting performance-rating errors in oral examinations. *Evaluation and the Health Professions* , 14 (1), 100–122.
- Raymond, M. R., and Viswesvaran, C. (1993). Least-squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement*, 30, 253-268.
- Redder, J. (2003). *Reliability: Rater's cognitive reasoning and decision making process* (Unpublished master's thesis). Portland State University, Portland, OR
- Reddy, Y. M., and H. Andrade. (2010). "A Review of Rubric Use in Higher Education." *Assessment & Evaluation in Higher Education* 35 (4): 435–448. doi:10.1080/02602930902862859.

- Reddy, Malini Y. (2011). "Design and Development of Rubrics to Improve Assessment Outcomes: A Pilot Study in a Master's Level Business Program in India." *Quality Assurance in Education* 19 (1): 84–104
- Reed, S. K., Dempster, A., & Ettinger, M. (1985). *Usefulness of analogous solutions for solving algebra word problems*. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11, 106–125.
- Reeves, T.C. & Okey, J.R. (1996). *Alternative assessment for constructivist learning environments*. In B.G. WILSON (Ed.), *Constructivist learning environments: Case studies in instructional design* (pp. 191-202). Englewood Cliffs, NJ: Educational Technology Publications
- Resnick, L. B., & Resnick, D. P. (1982). *Assessing the thinking curriculum: New tools for educational reform*. In B. G. Gifford & M. C. O'Conner (Eds.), *Changing assessment: Alternative views of aptitude, achievement and instruction* (pp. 37-55), Boston: Kluwer Academic.
- Resnick, L. B., & Resnick, D. P. (1992). *Assessing the thinking curriculum: New tools for educational reform*. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer
- Richland, L., Holyoak, K., Stigler, J. (2004), Analogy Use in Eighth-Grade Mathematics Classroom. *Cognition and Instruction*, 22 (1), p. 37-60, [http://reasoninglab.psych.ucla.edu/KH%20pdfs/Richand\\_etal.2004.pdf](http://reasoninglab.psych.ucla.edu/KH%20pdfs/Richand_etal.2004.pdf)
- R Yudha, P Marsukan. (2014). Pengembangan Instrumen Asesmen Otentik Unjuk Kerja Materi Bangun Ruang Di SMP. *Journal Education Research and Evaluation* 14 (2), 63-67.
- Rochmad. (2012). Desain Model Pengembangan Perangkat Pembelajaran Matematika. *Jurnal Kreano FMIPA UNNES*, vol.3 No.1, Hlm. 59-72 .
- Rockman, I. (2002). Strengthening connections between information literacy, general education and assessment efforts. *Library Trends* 51(2), pp. 185-198.
- Rost, J. (1988). Rating scale analysis with latent class models. *Psychometrika*, 53, 327-348.
- Rudner, L.M., and Schafer, W.D., (2002). *What Teachers Need to Know About Assessment*. Washington DC : National Education Association.

- Rulon, Michael (2002). Authenticity: The key to standards-based assessment Classroom. *Journal classroom leadership*. Vol 5 (9)
- Saad, N S. 2008. *Teaching Mathematics in Secondary Schools : Theories and Practices*. Perak : Universiti Pendidikan Sultan Idris.
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88(2), 413–428.
- Sadler, Philip M., and Eddie Good. (2006). “The Impact of Self-and Peer-grading on Student Learning.” *Educational Assessment* 11 (1): 1–31
- Sadler, D. R. (2009a). “Indeterminacy in the Use of Preset Criteria for Assessment and Grading.” *Assessment & Evaluation in Higher Education* 34 (2): 159–179. doi:10.1080/02602930801956059
- Sadler, D. R. (2009b). “*Transforming Holistic Assessment and Grading into a Vehicle for Complex Learning.*” In *Assessment, Learning and Judgement in Higher Education*, edited by G. Joughin, 1–19. Dordrecht: Springer.
- Sa'dijah, C. (2009). Asesmen Kinerja dalam Pembelajaran Matematika. *Jurnal pendidikan inovatif*, 4(2): 92-95. Tersedia di <http://jurnaljpi.files.wordpress.com/2009/09/vol-4-no-2-cholis-sadjah.pdf> diakses 7-2-2019.
- Sa'dijah, Cholis & Sukoriyanto. (2015). *Asesmen Pembelajaran Matematika*. Malang: UM Press
- Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment*, 11 (1), 1–31.
- Sarwiji Suwandi. (2011). *Model-model Asesmen dalam Pembelajaran*. Surakarta: Yuma Pustaka
- Schreiber, Lisa M., Gregory D. Paul, and Lisa R. Shibley. (2012). “The Development and Test of the Public Speaking Competence Rubric.” *Communication Education* 61 (3):205–233.
- Scullen, S. E. (1999). Using confirmatory factor analysis of correlated uniquenesses to estimate method variance in multitrait-multimethod matrices. *Organizational Research Methods*, 2, 275-292.

- Scullen, S. E., Mount, M. K., and Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, 85, 956-970.
- Sharef, N. M., Hamdan, H., & Madzin, H. (2014). Innovation-enhanced rubrics assessment for final year projects. *Global Journal of Engineering Education*, Volume 16, Number 3, 129–135.
- Shavelson, R. J., Webb, N. M., and Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, 44, 922-932.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage Publications.
- Shrout, P.E., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420-428.
- Silvestri, L, & Oescher, J. (2006). Using rubrics to increase the reliability of assessment in health classes. *International Electronic Journal of Health Education*, 9, 25–30
- Smith, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement*, 2(3), 281-311.
- Soeprijanto (2010). *Pengukuran Kinerja Guru Praktik Kejuruan*. Jakarta. CV. Tursina
- Spandel, V. (2006). *In defense of rubric*. English Journal, 96(1), 19–22.
- Sri Widiastuti, I.A.Md. (2016). EFL Teachers' Beliefs and Practices of Formative Assessment to Promote Active Learning. *The ASIAN EFL Journal*. Volume 3.
- Sri Widiastuti, I.A.Md. (2017). Teachers Understanding of Formative Assessment. *Jurnal Bahasa dan Seni*. Vol 45, No 1 Juni 2017.
- Stellmack, M. A., Konheim-Kalkstein, Y. L., Manor, J. E., Massey, A. R., & Schmitz, J. A. P.(2009). An assessment of reliability and validity of a rubric for grading APA-style introductions. *Teaching of Psychology*, 36(2), 102–107 <http://dx.doi.org/10.1080/00986280902739776>.
- Steinberg, L. S., Mislevy, R. J., Almond, R. G., Baird, A. B., Cahallan, C., Dibello, L. V., ...Kindfield, A. C. (2003). *Introduction to the biomass*

*project: An illustration of evidence-centered assessment design and delivery capability (CSE Technical Report 609).* Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Sternberg, Robert J. (2007). Assessing what matters. *Educational Leadership*, 65 (4) 20- 26.

Stephanie L. Knight, Scott P. McDonald. (2014). Professional Development and Practices of Teacher Educators. *Journal of Teacher Education*, vol. 65, 4: pp. 268-270., First Published August 5, 2014.

Steven Athanases. (1994). Teachers' Reports of the Effects of preparing portfolios of Literacy instruction. *Elementary school Journal* 94 (4): 421-439.

Stiggins, R. J. (1994). *Student-centered classroom assessment*. Merrill. New York.

Stiggins,R. J., & Bridgeford,N. J.(1985). The ecology of classroom assessment. *Journal of Educational Measurement*, 22,27 1-286.

Stiggins, R.J. (2001). *Student-involved classroom assessment*. 3rd ed. Upper Saddle River, NJ:Prentice-Hall.

Strijbos, J. W., Ochoa, T. A., Sluijsmans, D. M. A., Segers, M. S. R., & Tillema, H. H. (2009).*Fostering interactivity through formative peer assessment in (web-based)collaborative learning environments*. In C. Mourlas, N. Tsianos, & P. Germanakos(Eds.), *Cognitive and emotional processes in web-based education: Integrating humanfactors and personalization* (pp. 375–395). Hershey, PA: IGI Global.

Subali, Bambang. (2010a). Bias Item Tes Keterampilan Proses Sains Pola Divergen dan Modifikasinya sebagai Tes Kreativitas. *Jurnal Penelitian dan Evaluasi Pendidikan*, 14(2): 309-334.

Sudaryono. (2012). *Dasar-Dasar Evaluasi Pembelajaran*. Graha Ilmu. Yogyakarta.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and manyfacet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 9(3), 239– 61

Sujaya. (2013). Pengaruh Model Pembelajaran Asesmen Autentik Terhadap Hasil Belajar Matematika Dengan Kovariabel Motivasi Berprestasi (Eksperimen

- Pada Siswa Kelas V Sd Negeri 1 Gianyar ). (Online).  
[http://pasca.undiksha.ac.id/ejournal/index.php/jurnal\\_ep/article/view/688](http://pasca.undiksha.ac.id/ejournal/index.php/jurnal_ep/article/view/688). Diakses 12 Januari 2019
- Sukardjo dan Lis Permana Sari. (2009). *Metodologi Penelitian Pendidikan Kimia*. Yogyakarta: FMIPA UNY.
- Sukestiyarno. (2004). *Penerapan Strategi Berbasis Media dan Tehnologi dalam Mengajarkan Materi Matematika Perdana Sebagai Implementasi Kurikulum Berbasis Kompetensi*. Laporan Penelitian Due Like UNNES
- Sugiyono. (2009). *Metode Penelitian Pendidikan (Pendekatan Kuantitatif, Kualitatif, dan R&D)*. Bandung: Alfabeta.
- Suharjana, Agus. (2009). *Mengenal Bangun Ruang dan Sifat-Sifatnya di Sekolah Dasar*. Yogyakarta: Pusat Pengembangan dan Pemberdayaan Pendidik dan Tenaga Kependidikan (PPPPTK Matematika)
- Sumaryatun, Rusilowati, A., & Nugroho, S. E .(2016). Pengembangan instrument penilaian autentik kurikulum 2013 berbasis literasi sains pada materi bioteknologi. *Journal of Primary Education*.doi:p-ISSN 2252-6404
- Sunarti dan Rahmawati. (2014). *Penilaian Dalam Kurikulum 2013*. Penerbit ANDI. Yogyakarta.
- Suparsa, I. N., Mantra, I. B. N., & Widiastuti, I. A. M. S. (2017). Developing Learning Methods of Indonesian as a Foreign Language. *International Journal of Social Sciences and Humanities (IJSSH)*, 1(2),51-57.
- Susila, I. K. (2012). *Pengembangan Instrumen Penilaian Unjuk kerja (PerformanceAssessment) Laboratorium pada Mata Pelajaran Fisika sesuai Kurikulum TingkatSatuan Pendidikan SMA Kelas X di kabupaten Gianyar*. Tesis. Denpasar: Pascasarjana Universitas Pendidikan Ganesha
- Susilaningsih E. (2018). Development of performance assessment instrument based contextual learning for measuring students laboratory skills.*IOP Conf. Ser.: Mater. Sci. Eng.* 349012018.doi:10.1088/1757-899X/349/1/012018
- Suwaji, U. T. (2018). *Permasalahan pembelajaran geometri ruang SMP dan alternatif pemecahannya*. Yogyakarta: PPPPTK
- Terwilliger, J. S. (1998). *Rejoinder: response to wiggins and newmann*. Educational Researcher, 27(6), 22-23.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Tierney, R., and M. Simon. (2004). "What's Still Wrong with Rubrics: Focusing on the Consistency of Performance Criteria across Scale Levels." *Practical Assessment, Research & Evaluation* 9 (2): 1–10

Tileston, Donna Walker (2005). *10 Best teaching practices, How brain research, learning styles, and standards define teaching competencies*, Second Edition. Thousand Oaks, CA: Corwin Press.

Tim Penyusun.(2014). *Permendikbud Nomor 104 Tahun 2014 tentang Penilaian Hasil Belajar oleh Pendidik pada Pendidikan Dasar dan Menengah*. Kemdikbud. Jakarta.

Timmerman, B. E. C., D. C. Strickland, R. L. Johnson, and J. R. Payne. (2010). "Development of a 'Universal' Rubric for Assessing Undergraduates' Scientific Reasoning Skills Using Scientific Writing." *Assessment & Evaluation in Higher Education* 36 (5): 509–547. doi:10.1080/02602930903540991.

Timmerman, Briana E. Crotwell, Denise C. Strickland, Robert L. Johnson, and John R. Payne. (2011). "Development of a 'Universal' Rubric for Assessing Undergraduates' Scientific Reasoning Skills Using Scientific Writing." *Assessment & Evaluation in Higher Education* 36 (5): 509–547. doi:10.1080/02602930903540991.

TIMSS. (1995). [Online]. *Performance assessment in IEA's third international mathematics and science study*. <http://www.timss.bc.edu/timss1995i/PReport.html>. (Accessed on 10 Februari 2019).

TIMSS. (1999). [Online]. *TIMSS 1999 international mathematics report findings from IEA's repeat of the third international mathematics and science study at the eighth grade*. [http://www.timss.bc.edu/timss1999i/math\\_achievement\\_report.html](http://www.timss.bc.edu/timss1999i/math_achievement_report.html). (Accessed on 10 Februari 2019).

TIMSS. (2003). [Online]. *TIMSS 2003 international mathematics report*. <http://timss.bc.edu/timss2003i/mathD.html>. (Accessed on 10 Februari 2019).

TIMSS. (2007). [Online]. *TIMSS 2007 international mathematics report findings from IEA's repeat of the third international mathematics and science study at the eighth grade.* [http://www.timss.bc.edu/timss2007i/math\\_achievement\\_report.html](http://www.timss.bc.edu/timss2007i/math_achievement_report.html). (Accessed on 10 Februari 2019).

Toulmin, Stephen (1958) : *The Uses of Argument*, Cambridge University Press, Cambridge.

Torgerson, W. F. (1961). Scaling and test theory. *Annual Review of Psychology*, 12 , 51–70.

Torgerson, W. S. (1958). *Theory and methods of scaling* . New York: Wiley.

Torrance, H. (1995). *Introduction.* In H. TORRANCE (Ed.), *Evaluating Authentic Assessment: Problems And Possibilities In New Approaches To Assessment* (pp. 1-8). Buckingham: Open University Press.

Torrance, H. (2007). “Assessment as Learning? How the Use of Explicit Learning Objectives, Assessment Criteria and Feedback in Post-secondary Education and Training can come to Dominate Learning.” *Assessment in Education: Principles, Policy & Practice* 14 (3):281–294. doi:10.1080/09695940701591867

Traub, R. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, 16 (10), 8–13.

Trianto. (2007). *Model Pembelajaran Terpadu dalam Teori dan Praktek*. Surabaya: Pustaka Ilmu

Tsai, C.-C., & Liang, J.-C. (2009). The development of science activities via on-line peer assessment: The role of scientific epistemological views. *Instructional Science*, 37, 293–310.

Tseng, S.-C., & Tsai, C.-C. (2007). On-line peer assessment and the role for of the peer feedback: A study of high school computer course. *Computers and Education*, 49, 1161–1174.

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory* . New York: Springer.

Van Tassel-Baska, J. (2014). Performance-Based Assessment The Road to Authentic Learning for the Gifted. *Gifted Child Today*, 37(1), 41-47.

- Van De Walle, J (2008). *Matematika sekolah dasar dan menengah Jilid 2.* (Terjemahan Suyono). Jakarta: Erlangga.
- Vandenberg, Amy, Matthew Stollak, Linda McKeag, and Doug Obermann. (2010). "GPS in the Classroom: Using Rubrics to Increase Student Achievement." *Research in Higher Education Journal* 9: 1–10.
- Velde, C. & Cooper, (2000). Students perspectives of workplace learning and training in vocational education. *Education and Training*. 42(2):83-92.
- Wahyudi, A. (2014). *Profil Dasar Keterampilan Dasar Geometri Siswa SD dalam Menyelesaikan Soal bangun Datar Ditinjau dari Gaya kognitif.* Tesis tidak diterbitkan. Surabaya: Program Pasca Sarjana Universitas Negeri Surabaya.
- Wahyuni, L.G.E. (2013). Authenticity of Teachers' Made Assessment and Its Contribution to Students' English Achievement. *Jurnal Pendidikan dan pengajaran Undiksha*, 46 (2): p. 1 82-1 91 .
- Wald, Hedy S., Jeffrey M. Borkan, Julie Scott Taylor, David Anthony, and Shmuel P. Reis.(2012). "Fostering and Evaluating Reflective Capacity in Medical Education: Developing the REFLECT Rubric for Assessing Reflective Writing." *Academic Medicine: Journal of the Association of American Medical Colleges* 87 (1): 41–50.doi:10.1097/ACM.0b013e31823b55fa
- Walle, J.A van De. (2001). *Geometric Thinking and Geometric Concept. In Elementar yand Middle School Mathematics.Teaching Developmentally 4th. Ed.* Boston: Pearson Education.
- Wang, W. (1997). *Estimating rater severity with multilevel and multidimensional item response modeling.* Taipei, Taiwan: Taiwan National Science Council. (ERIC Document Reproduction Service No. ED 408 340).
- Wang, W., Wilson, M. R., and Adams, R. J. (1997). Rasch models for multidimensionality between and within items. In M. Wilson, G. Engelhard, Jr., and K. Draney (Eds.), *Objective measurement: Theory into practice:* Vol. 4 (pp. 139-156). Greenwich, CT: Ablex.
- Wedgege, T. (2013). *Workers mathematical competences as a study object: Implications of general and subjective approaches.* Malmö. Faculty of Education and Society. Adults' mathematics: Working Papers, 2. Retrieved from site: <http://www.mah.se/Is/eng>

- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, UK: Palgrave Macmillan.
- Widoyoko, S. Eko Putro. (2014). *Evaluasi Program Pembelajaran: Panduan Praktis Bagi Pendidik dan Calon Pendidik*. Pustaka Pelajar, Yogyakarta
- Weigle, S. C. (1998). *Using FACETS to model rater training effects*. *Language Testing*, 15, 263–287.
- Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, 6, 145–178.
- Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.
- Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitraitmultimethod data. *Applied Psychological Measurement*, 9, 1-26.
- Wiggins, G. (1989). “A True Test: Toward More Authentic and Equitable Assessment”. *Phi Delta Kappa International*. 70(9): 703 – 713.
- Wiggins, G. (1989a). Teaching to the (authentic test). *Educational Leadership*, 46(7), 41-47.
- Wiggins,G.(1993). Assessment to improve performance, not just monitor it: Assessment reform in the social sciences. *Social Science Record*, 30(2), 5-12.
- Wiggins, G. (1993). *Assessing Student Performance – Exploring the Purpose and Limits of Testing*. San Francisco, CA: Jossey-Bass.
- Wiggins, G. (1996). Practicing what we preach in designing authentic assessments. *Educational Leadership*, 18-25.
- Wiggins, G. 2011. “Kappan Classic: A TrueTest: Toward more Authentic and Equitable Assessment”. *Phi Delta Kappan*. April 2011 92 (7): 81-93 Bloomington.Diambil pada 27 Mei 2019, dari <http://proquest.umi.com/pqdweb>.
- Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10, 305–335.

- Wigglesworth, G. (1994). Patterns of rater behavior in te assessment of an oral interction test. *Australian Review of Applied Linguistics*, 17(2), 77–103.
- Williams, R.G., Klamen, D.A., & McGaghie, W.C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine*, 15, 270-292.
- Wilson, H. G. (1988). Parameter estimation for peer grading under incomplete design. *Educational and Psychological Measurement*, 48, 69-81.
- Wilson, M. R., and Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, 60, 181-198.
- Wilson, M. R., and Case, H. (2000). An examination of variation in rater severity over time: A study of rater drift. In M. Wilson, and G. Engelhard, Jr. (Eds.), Objective measurement: *Theory into practice*: Vol. 5 (pp. 113-134). Stamford, CT: Ablex.
- Wilson, M. R., and Hoskens, M. (1999, April). *The rater bundle model for constructed-response items: An example in the context of real-time feedback on rater effects*. Paper presented at the annual meeting of the American Educational Research Association, Montréal, Canada.
- Wiyono, Bambang, dan Sunarni. (2009). *Evaluasi Program Pendidikan dan Pembelajaran*. Malang: Fakultas Ilmu Pendidikan Universitas Negeri Malang.
- Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, 4, 83–106.
- Wolfe, E. W. (1998, April). *A two-parameter logistic rater model (2PLRM): Detecting rater harshness and centrality*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA
- Wolfe, E. W. (in press). Identifying rater effects using latent trait models. *Psychology Science*.
- Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, 15 , 465–492.

- Wolfe, E. W., Chiu, C. W. T., and Myford, C. M. (1999). *The manifestation of common rater effects in multi-faceted Rasch analyses* (MS #97-02). Princeton, NJ: Educational Testing Service, Center for Performance Assessment
- Wolfe, E. W. (2000). Equating and item banking with the Rasch model. *Journal of Applied Measurement*, 1, 409–434.
- Wolfe, Patricia (2001). *Brain matters translating research into classroom practice*. Alexandria, Virginia: Association for Supervision and Curriculum Development.
- Wolfe, E. W., Moulder, B. C., & Myford, C. M. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied Measurement*, 2, 256–280.
- Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, 46, 35–51.
- Wolfe, E. W., Myford, C. M., Engelhard, G., & Manalo, J. R. (2007). *Monitoring reader performance and DRIFT in the AP English Literature and Composition examination using benchmark essays* (College Board Research Report No. 2007-2). New York: College Board.
- Wolfe, E. W. (2008). RBF.sas (Rasch Bootstrap Fit): A SAS macro for estimating critical values for Rasch model fit statistics. *Applied Psychological Measurement*, 32, 585–586.
- Wolfe, E. W., & Dobria, L. (2008). *Applications of the multifaceted Rasch model*. In J. W. Osborne (Ed.), Best practices in quantitative methods (pp. 71–85). Los Angeles: Sage.
- Wolfe, E. W. (2009). Item and rater analysis of constructed response items via the multi-faceted Rasch model. *Journal of Applied Measurement*, 10, 335–347.
- Wu, M., Adams, R., and Wilson, M. (1997). *ConQuest* [Computer program]. Melbourne, Australia: Australian Council for Educational Research.
- Yohana PSR, M.Makhsuli, AJ Purwanto, YW Paskalis, Erfan Yudianto. (2006). *Pengembangan Media Pembelajaran Matematika berbantuan Komputer*. Makalah. Disampaikan dalam Pekan Ilmiah Mahasiswa Nasional: UMM

Yopp, By David, and Richard Rehberger. (2009). "A Curriculum Focus Intervention's Effectson Prealgebra Achievement." *Journal of Developmental Education* 33 (2): 28–30

Yudabakan, İ. (2011) . The Influence of Peer and Self-Assessment on Learning And Metacognitive Knowledge: Consequential Validity. *International Journal on New Trends in Education and Their Implications*, 2(4), 44-57.

Yusuf, A.M. (2015). *Asesmen dan Evaluasi Pendidikan, Pilar Penyedia Informasi dan Kegiatan Pengendalian Mutu Pendidikan*. Jakarta: Kencana.

Zacharis, N.T. (2010). Innovative Assessment for Learning Enhancement: Issues and Practices. *Contemporary Issues in Education Research. (Online)*, 3 (1):61 —70, (<http://files.eric.ed.gov/fulltext/EJ1072576.pdf>, diakses 31 Oktober 2018).

Zainul, A. (2001). *Alternative Assessment. Applied Approach Mengajar di Perguruan Tinggi*. Jakarta: Pusat Antar Universitas untuk Peningkatan dan Pengembangan Aktivitas Instruksional. Ditjen Dikti Depdiknas.

Zheng, Changlong, Lihai Fu, & Peng He. (2014). Development of an Instrument for Assessing the Effectiveness of Chemistry Classroom Teaching. *J Sci Educ Technol*, 23: 267-279.