Teacher evaluation through observation

van der Lans, Rikkert

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123–140.

Antoniou, P., Kyriakides, L., & Creemers, B. P. M. (2015). The Dynamic Integrated approach to teacher professional development: rationale and main characteristics. *Teacher development, 19*(4), 535-552. doi: 10.1080/13664530.2015.1079550

Bachman, L. F. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice*, *21*(3), 5-18.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear Mixed-Effects Models Using S4 Classes*. R package version 1.1-7. URL: http://CRAN.R-project.org/package=lme4.

Berliner, D. (2001). Learning about learning from expert teachers. *International Journal of Educational Research, 35,* 463–483.

Benton, S. L., & Cashin, W. E. (2012). *Student ratings of teaching: a summary of the research and literature.* (IDEA Paper No. 50). Retrieved March 3, 2015, from http://www.ntid.rit.edu/sites/default/files/academic_affairs/Sumry%20of%20Res%20%2350%20Benton%202012.pdf.

Borko, H. (2004). Professional Development and Teacher Learning: Mapping the Terrain, *Educational Researcher, 33*, 3-15. doi: 10.3102/0013189X033008003

Bill & Melinda Gates Foundation (2012). *Asking students about teaching: Student perception surveys and their implementation.* Retrieved March 3 from http://www.metproject.org/downloads/Asking_Students_Practitioner_Brief.pdf.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* London: Lawrence Erlbaum Associates.

Brennan, R. L. (2001). *Generalizability theory: Statistics for social science and public policy.* New York: Springer-Verlag.

Brennan, R. L. (2004). *Some perspectives on inconsistencies among measurement models.* (CASMA Research Report No. 10). Retrieved March 9, 2015, from http://www.uiowa.edu/~casma/NSF-casma-rpt.pdf.

Briggs, D. C., & Wislon, M. (2007). Generalizability in Item response theory. *Journal of Educational Measurement, 44*, 131 – 155.

Browne, M. W. (1992). Circumplex models for correlation matrices. *Psychometrika, 57,* 469-497.

Cai, L., Thissen, D., & Du Toit, S. (2005–2013). IRTPRO (Version 2.1) [Computer software]. Lincolnwood, IL: Scientific Software International.

Centra, J. A. (1975). Colleagues as raters of classroom instruction. *The Journal of Higher Education, 46*(3)*,* 327-337. doi: 10.2307/1980806

Chen, W-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.

Cheung, W. M., & Wong, W. M. (2013). Does lesson study work? A systematic review on the effects of lesson study and learning study on teachers and students. *International Journal for Lesson and Learning Studies, 3*(2), 137-149. DOI 10.1108/IJLLS-05-2013-0024

Choi, J. (2013). *Advances in combining generalizability theory and item response theory.* Doctoral dissertation, University of California, Berkeley.

Conway, P. F., & Clark, C. M. (2003). The journey inward and outward: A re-examination of Fuller's concerns-based model of teacher development. *Teaching and Teacher Education, 19*, 465–482.

CPS, (2012). *Ziet u het verschil? Ook in het voortgezet onderwijs is differentiëren essentieel* [*Do you see the difference? The relevance of differentiation in secondary education*] [special issue]. Didactief, 42(8).

Creemers, B. P. M., & Kyriakides, L. (2005) Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement, 17*(3), 347-366. doi: 10.1080/09243450600697242

Cromley, J. G., Perez, T. C., Fitzhugh, S. L., Newcombe, N. S., Wills, T. W., & Tanaka, J. C. (2013). Improving students' diagram comprehension with classroom instruction. *Journal of experimental education, 81,* 511-537. doi: 10.1080/00220973.2012.745465

Cronbach, L., J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-333.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp 3-17). New Jersey, USA: Lawrence Erlbaum Associates Inc.

Cronbach, L. J. (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and psychological measurement, 64,* 391-418. doi: 10.1177/0013164404266386

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

Cronbach, L. J., Gleser, C. G., Rajaratnam, N., & Nanda, H. (1972). *The dependability of behavioral measurements.* New York, USA: Wiley.

Dall'Alba, G., & Sandberg, J. (2006). Unveiling Professional Development: A Critical Review of Stage Models. *Review of educational research, 76*(3), 383-412.

Danielson, C. (2013). *The framework for teaching: Evaluation instrument.* Princeton, NJ: The Danielson Group.

Darling-Hammond, L. (2013). *Getting teacher evaluation right. What really matters for effectiveness and improvement.* New York, USA: Teachers College Press

Darling-Hammond, L., Amrein-Beardsley, A., Heartel, E., & Rothstein J. (2012). Evaluation teacher evaluation. *Phi Delta Kappan, 93,* 8–15.

Day, C., Sammons, P., Stobart, G., Kingston, A., & Gu, Q. (2007). *Teachers matter: Connecting lives, work and effectiveness.* Maidenhead, UK: Open University Press.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Abe, H., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39,* 1–25.

De Jong, R., & Westerhof, K.J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research 4,* 51–85.

DeMars, C. (2010). *Item response theory: Understanding statistics measurement.* New York: Oxford University Press.

DfEE (2012). *Teacher appraisal and capability. A model policy for schools.* Retrieved June 30, 2015, from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/282598/Teacher_appraisal_and_capability.pdf

Doran, H., Bates, D., Blies, P., & Dowling, M. (2007). Estimating the multilevel Rasch model with the lme4 Package. *Journal of Statistical Software, 20*, 1-18.

Doyle, W. (1983). Academic work. *Review of Educational Research, 53*(2), 159-199.

Doyle, W. (2006). Ecological approaches to classroom management. In: C.M. Evertson & C.S. Weinstein (Eds), *Handbook of classroom management*: *research, practice, and contemporary issues* (p. 97- 125). New York: Erlbaum.

Doyle, W. (2009). Situated Practice: A Reflection on Person-Centered Classroom Management. *Theory Into Practice, 48*(2), 156-159, doi: 10.1080/00405840902776525

Ebbens, S., & Ettekoven, S. (2005). *Effectief leren: Basisboek* [*Effective learning: Basic handbook*]. Groningen, The Netherlands: Wolters-Noordhoff.

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education, 30*(2), 137-194.

Firestone, W. A. (2014). Teacher evaluation policy and conflicting theories of motivation. *Educational Researcher 43,* 100-107.

Fox, J-P. (2010). *Bayesian Item Response Modeling. Theory and Applications.* New York: USA, Springer

French-Lazovik, G. (1981). Documentary evidence in the evaluation of teaching. In J. Millman (Ed.), *Handbook of Teacher Evaluation*. Beverly Hills, CA: Sage Publications.

Fuller, F. (1969). Concerns of teachers: A developmental conceptualization. *American Educational Research Journal, 6*, 207–226.

Gelman A., Su Y-S, Yajima M, Hill J, Pittau M. G., Kerman J., et al. *Arm: Data analysis using regression and multilevel/ hierarchical models*. R package version 1.8-6 2015: URL: https://cran.r-project.org/web/packages/arm/arm.pdf.

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of teacher performance. *Economica, 80*, 589–612. doi:10.1111/ecca.12002

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher, 43,* 293 – 303. DOI:10.3102/0013189X14544542

Guilleux, A., Blanchin, M., Hardouin, J-B., Sébille, V. (2014). Power and Sample Size Determination in the Rasch Model: Evaluation of the Robustness of a Numerical Method to Non-Normality of the Latent Trait. *Plus One, 9*(1), 1-7.

Guttman, L. L. (1954). A new approach to factor analysis: the radex. In Lazersfeld, Paul F. (Ed.), *Mathematical thinking in the social sciences*. Glencoe, Illinois: The Free Press

Guttman, L. L. (1977). What is not what in statistics. *Journal of the Royal Statistical Society, 26*(2), 81-107.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Abingdon, Oxon, UK: Routledge.

Haberman, S. J. (2008) When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*(2), 204-229. doi: 10.3102/1076998607302636

Hanushek, E. A. (2007) . The single salary schedule and other issues of teacher pay. *Peabody Journal of Education, 82*, 574-586. doi: 10.1080/01619560701602975

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30*(3), 466-479. doi: 10.1016/j.econedurev.2010.12.006

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., Brown, J. L., Cappella, E., Atkins, M., Rivers, S. E., Brackett, M. A., & Hamagami, A. (2013). Teaching through interactions: testing a developmental framework of teaching effectiveness in over 4,000 classrooms. *The Elementary School Journal, 113,* 461-487.

Hazi, H. M., & Rucinsky, D. A. (2009). Teacher Evaluation as a Policy Target for Improved Student Learning: A Fifty-State Review of Statute and Regulatory Action since NCLB. *Educational Policy Analysis Archives, 17*(5), 1-22.

Helms-Lorenz, M., Van de Grift, W. J. C. M., & Maulana, R. (2016). Longitudinal effects of induction on teaching skills and attrition rates of beginning teachers. *School Effectiveness and School Improvement, 27*(2), 178-204. Doi: 10.1080/09243453.2015.1035731

Hill, H. C., Besiegel, M., & Jacob, R. (2013). Professional development research: Consensus, crossroads and challenges. *Educational Researcher, 42,* 476-487. doi: 10.3102/0013189X13512674

Hill, H., Charalambous, C. Y., & Kraft, M. A. (2012). When interrater-reliability is not enough: Teacher observation systems and a case for the generalizability theory. *Educational Researcher 41,* 56–64. doi: 10.3102/0013189X12437203.

Hill, H., Kapitula, L., & Umland, K. (2011). A validity argument to evaluating teacher value added scores. *American Educational Research Journal 48*, 797–831.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel.* Seattle, WA: Bill & Melinda Gates Foundation.

Howard, G. S., Conway, C. G., & Maxwell, S. E. (1985). Construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology, 77*(2), 187-196.

Hoxby, C. M. (2002). *The cost of accountability* (No. w8855). National Bureau of Economic Research.

Hox, J. (2010). Multilevel analysis: Techniques and applications (2nd edition). New York, NY: Routledge.

Hu, L., & Bentler, M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi: 10.1080/10705519909540118

Huberman, M. (1993). *The lives of teachers.* New York, NY: Teachers College Press.

Inspectie van het Onderwijs (2009). *International Comparative Analysis of Learning and Teaching in Math Lessons in Several European Countries.* De Meern, Inspectie van het Onderwijs.

Inspectie van het Onderwijs (2016). *De staat van het onderwijs 2014-2015* [*The state of education in the Netherlands 2014-2015*]. De Meern, Inspectie van het Onderwijs.

Inspectie van het Onderwijs (2015). *De staat van het onderwijs 2013-2014* [*The state of education in the Netherlands 2013-2014*]. De Meern, Inspectie van het Onderwijs.

Isoré, M. (2009). *Teacher Evaluation: Current Practices in OECD Countries and a Literature Review.* OECD Education Working Papers, No. 23. OECD Publishing (NJ1).

Jöreskog, K. G. (1970). Estimation and testing of simplex models. *The British Journal of Mathematical and Statistical Psychology, 23*(2), 121-145.

Kagan, D. M. (1992). Professional growth among pre-service and beginning teachers. *Review of Educational Research, 62*, 129–169.

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

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1-73. doi: 10.1111/jedm.12000

Kane, T. J., Staiger, D. O., McCaffrey, D., Cantrell, S., Archer, J., Buhayar, S., Kerr, K., Kawakita, T., & Parker, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation.

Kennedy, M. (2010). Attribution error and the quest for teacher quality. *Educational researcher, 39*, 591-598. doi: 10.3102/0013189X10390804

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (Third edition). New York: Guilford Press.

Koller, I. & Hatzinger R. (2013). Nonparametric tests for the Rasch model: explanation, development, and application of quasi-exact tests for small samples. *Interstat, 11,* 1-16.

Kyriakides, L. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education, 36*, 143–152.

Kyriakides, L., Creemers, B. P. M., & Antaniou, P. (2009). Teacher behavior and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education, 25,* 12–23.

Louws, M. (2016). *Professional learning: What teachers want to learn* (Doctoral Dissertation). Leiden, the Netherlands: ICLON.

Mainhard, T. M., Brekelmans, M., Den Brok, P., & Wubbels, T. (2011). The development of the classroom social climate during the first months of the school year. *Contemporary Educational Psychology, 36*(3), 190-200. doi:10.1016/j.cedpsych.2010.06.002

Mair, P., & Hatzinger, R. (2007). Extended Rasch modelling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*, 1–20.

Marsh, H. D. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Dordrecht, The Netherlands: Springer.

Marzano, R.J. (2003). *What works in schools: Translating research into action.* Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J. (2012). The two purposes of teacher evaluation. *Educational Leadership, 70,* 14-19.

Marzano, R. J., & Toth, M. D. (2013). *Teacher evaluation that makes a difference. A new model for teacher growth and student achievement.* Alexandria, Virginia: ASCD

Maualan, R., & Helms-Lorenz, M. (2016). Observations and student perceptions of the quality of preservice teachers' teaching behaviour: construct representation and

predictive quality. *Learning environments research, 19*(3), 335-357. doi:10.1007/s10984-016-9215-8

Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2015). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement, 26*(2), 169-194.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*(4)*,* 311–314.

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14*(3), 283-298.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching.* Seattle, WA: Bill & Melinda Gates Foundation.

Mourshed, M., Chijioke C., & Barber, M. (2010). *How the world's most improved school systems keep getting better.* London: McKinsey Company.

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation: An International Journal on Theory and Practice 12,* 53–74.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art—teacher effectiveness and professional learning, *School Effectiveness and School Improvement, 25*(2), 231–256, doi: 10.1080/09243453.2014.885451

Murray, H. G. (1983). Low-inference classroom teaching and student ratings of college teaching effectiveness. *Journal of Educational Psychology, 75*(1), 138-149.

Murillo, F. J. (2007). *Evaluación del desempeño docente y carrera profesional docente. Un estudio comparado entre 50 países de América y Europa.* Santiago de Chile: OREALC/UNESCO.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles: Muthén and Muthén.

NCTQ (2013). *Connect the dots: Using evaluations of teaching effectiveness to inform policy and practice.* Washington, DC: National Council on Teacher Quality.

Nusche, D., Braun, H., Halász, G., & Santiago, P. (2014). *OECD Reviews of Evaluation and Assessment in Education: Netherlands 2014.* OECD Reviews of Evaluation and Assessment in Education, OECD Publishing.

http://dx.doi.org/10.1787/9789264211940-en

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

OCW [Dutch Ministry of Education, Culture, and Science] (2013a). *Peer review in de praktijk* [*Peer review in practice*]. Rotterdam: VOION.

OCW [Dutch Ministry of Education, Culture, and Science] (2013b). *Begeleiding van beginnende leraren in het beroep* [*Induction of inexperienced teachers*]. The Hague, The Netherlands: OCW: http://www.leroweb.nl/cms/wp-content/uploads/2013/11/Begeleiding-beginnende-leraren.pdf

OECD (2016), *Netherlands 2016: Foundations for the Future*, OECD Publishing, Paris. doi: http://dx.doi.org/10.1787/9789264257658-en

Overdiep I. (2016). *Wijzer over Zien en Kijken: Inventarisatie observatie instrumenten in het PO* [*learning from observation: a review of observation instruments applied in Primary Education*]. Utrecht, The Netherlands: PO-raad. https://www.poraad.nl/files/werkgeverszaken/wijzer_over_zien_en_kijken.pdf

Patrick, H., & Mantzicopolous, P. (2016). Is effective teaching stable? *Journal of Experimental Education, 84,* 23-47. doi: 10.1080/00220973.2014.952398

Patz, R. P., Jucker B. W., Johnson, M. S., & 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. doi: 10.3102/10769986027004341

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practice.* Thousand Oaks, CA: Corwin Press.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: standardized observation can leverage capacity. *Educational researcher, 38*(2), 109-119. doi: 10.3102/0013189X09332374

Ponocny, I. (2001). Non-parametric goodness-of-fit tests for the Rasch model. *Psychometrika 66*, 437–460.

Popham, (1988). The dysfunctional marriage of formative and summative teacher evaluation. *Journal of Personnel Evaluation in Education, 1,* 269 – 273.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2006). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement, 30,* 1–12. doi: 10.1177/0146621606291569.

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment and Evaluation in Higher Education, 30*, 387–415. doi: 10.1080/02602930500099193.

Richardson, V., & Placier, A. (2001). Teacher change. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed.). Washington, DC: American Educational Research Association.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*, 1–25.

Schafer, E., Stringfield, S., & Wolfe, D. (1992). Two-year effects of a sustained beginning teacher induction program on classroom interactions. *Journal of Teacher Education, 43,* 203–214.

Shepard, L. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 405–450). Washington, DC: American Educational Research association.

Scriven, M. (1981). Summative teacher evaluation. In J. Millman (Ed.), *Handbook of Teacher Evaluation*. Beverly Hills, CA: Sage Publications.

Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of teaching behavior. *Review of Educational Research, 46,* 553-611.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer*. Thousand Oaks, California: Sage Publications, Inc

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1-23.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72-101.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271-295.

Steffy, B. E., & Wolfe, M. P. (2001). A life-cycle model for career teachers. *Kappa Delta Pi Record 38*, 16–19. doi: 10.1080/00228958.2001.10518508.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Difference, 42*, 893-898.

Sternberg, R., J. & Horvath, J. A. (1995). A prototype view of expert teaching. *Educational Researcher, 24*, 9–17.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55(2), 293-325.

Strong, M. (2011). *The high qualified teacher. What is teacher quality and how do we measure it?* NY, NY: Teachers College press

Teitler, P. (2013). *Lessen in orde* [*Lessons in order*]. Bussum, The Netherlands: Uitgeverij Couthino

Tendeiro, J. N. (2014). Package 'PerFit' (published online). In R. Cran (Ed.), *The comprehensive R network.* retrieved from: http://cran.r-project.org/web/packages/PerFit/PerFit.pdf.

Thorndike, E. L. (1918). Individual differences. *Psychological Bulletin*, *15*, 148–159.

Timmerman, M. E., Lorenzo-Seva, U. & Ceulemans, E. (in press). The number of factors problem. in P. Irwing, T. Booth, & D.J. Hughes. (eds.; in press). *The Wiley Handbook of Psychometric Testing*, John Wiley & Sons, Chichester, UK.

Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public schools.* Washington DC: Education Sector.

U.S. Department of Education (2009). *Race to the Top Program: Executive Summary*. Washington, USA.

Vale, C. D. (1986). Linking Item parameters onto a common scale. *Applied Psychological Measurement, 10*(4), 333-344.

Van Veen, K., Zwart, R., Meirink, J., & Verloop, N. (2010). *Professionele ontwikkeling van leraren* [*Professional development of teachers*]. Leiden, The Netherlands: ICLON: Expertisecentrum leren van docenten.

Van de Grift, W. (1990). Het onderzoek naar effectieve scholen. *Pedagogische Studiën*, *67* (10) 462-463

Van de Grift, W. J. C. M. (2007). Quality of teaching in four European countries: a review of the literature and application of an assessment instrument. *Educational Research, 49*(2), 127-152. doi: 10.1080/00131880701369651

Van de Grift, W. J. C. M. (2013). *Van zwak naar sterk. De aanpak van zwakke en zeer zwakke scholen voor voortgezet onderwijs in het noorden van het land door*

observatie van en feedback voor leraren. Drachten: The Netherlands. https://www.rug.nl/staff/w.j.c.m.van.de.grift/vanzwaknaarsterkdrachten.pdf

Van de Grift, W. J. C. M. (2014). Measuring teaching quality in several European countries. *School effectiveness and school improvement, 25*(3), 295–311 doi: 10.1080/09243453.2013.794845

Van de Grift, W. J. C. M. & J.F. Lam (1998). Het didactisch handelen in het basisonderwijs. [Teachers' instructions in primary education.] *Tijdschrift voor Onderwijsresearch 23*(3), 224-241.

Van de Grift, W. J. C. M., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation, 43*, 150–159 doi: 10.1016/j.stueduc.2014.09.003

Van de Grift, W. J. C. M., Van der Wal, M., & Torenbeek, M. (2011). Ontwikkeling in de pedagogische didactische vaardigheid van leraren in het basisonderwijs. [Primary teachers' development of pedagogical didactical skill.] *Pedagogische Studiën, 88*, 416–432.

Van der Ark, A., L. (2007). Mokken Scale Analysis in R. *Journal of Statistical Software, 20*(11), 1-19. doi: 10.18637/jss.v020.i11

Van der Lans, R. M., Van de Grift, W. J., & Van Veen, K. (2015). Developing a Teacher Evaluation Instrument to Provide Formative Feedback Using Student Ratings of Teaching Acts. *Educational Measurement: Issues and Practice, 34*(3), 18-27.

Van der Lans, R. M., Van de Grift, W. J. C. M., & Van Veen (2017). Developing an instrument for teacher feedback: Using the Rasch model to explore teachers' development of effective teaching strategies and behaviors. Journal of Experimental Education (first online publication). doi: 10.1080/00220973.2016.1268086

Van der Lans, R. M., Van de Grift, W. J. C. M., Van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluative decisions using classroom observation. *Studies in Educational Evaluation, 50*, 88-95.

Van Veen, K. (2011). Het niveau en de kwalitateit van leraren in het basisonderwijs en voortgezet onderwijs: wat is het probleem? [the competence and quality of primary and secondary education teachers: what is the problem?] *Pedagogische studiën*, 433-441.

Van Veen, K., Zwart, R., Meirink, J., & Verloop, N. (2010). *Professionele ontwikkeling van leraren: Een reviewstudie naar effectieve kenmerken van professionaliseringsinterventies van leraren* [*Teacher professional development: A review study on effective characteristics of teacher professional development interventions*]. Leiden: ICLON / Expertisecentrum Leren van Docenten [ICLON / Expertise centre Teacher learning].

Van Schuur, W. H. (2011). *Ordinal Item response theory: Mokken scale analysis* (Vol 169). Thousand Oaks, California: Sage publications

Vitikka, E., Krokfors, L., & Hurmerinta, E. (2012). *The Finnish National Core Curriculum: Structure and development* (draft). Retrieved September 2016 from: http://curriculumredesign.org/wp-content/uploads/The-Finnish-National-Core-Curriculum_Vitikka-et-al.-2011.pdf

Von Davier, M., & Carstensen, C. H. (2007). *Multivariate and mixture distribution Rasch models: extensions and applications.* Springer: New York.

Weeks, J. P. (2010). plink: A R Package for linking mixed-format tests using IRT-Based methods. *Journal of Statistcal Software 35*(12), 1-33. doi: 10.18637/jss.v035.i12.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* New Teacher Project.

Winters, M. A., & Cowen, J. M. (2014). Who would stay, who would be dismissed? An empirical consideration of value-added teacher retention policies. *Educational Researcher, 42,* 330–337. DOI: 10.3102/0013189X13496145

Wubbels, T., & Brekelmans, M. (2005). Two decades of research on teacher-student relationships in class. *International Journal of Educational Research, 43,* 6-24. doi:10.1016/j.ijer.2006.03.003