D. Measurement.

What is inter-rater agreement and how can designers of teacher evaluation systems maximize it?

This synthesis of key research studies was written by:

Matthew Graham, Westat, November 2011

We are grateful to Anthony Milanowski, Westat, for his helpful comments and suggestions.

The work described in this paper was supported by the U.S. Department of Education through the Center for Educator Compensation Reform. The opinions expressed are those of the authors and do not necessarily reflect the view of the U.S. Department of Education, the Center for Educator Compensation Reform, or the institutional partners of the Center. Comments and suggestions are welcome.

The Center for Educator Compensation and Reform (CECR) was awarded to Westat — in partnership with Learning Point Associates, an affiliate of American Institutes for Research, Synergy Enterprises Inc., J. Koppich and Associates and the University of Wisconsin — by the U.S. Department of Education in October 2006.

The primary purpose of CECR is to support Teacher Incentive Fund (TIF) grantees in their implementation efforts through provision of sustained technical assistance and development and dissemination of timely resources. CECR also is charged with raising national awareness of alternative and effective strategies for educator compensation through a newsletter, a Web-based clearinghouse, and other outreach activities.

This work was originally produced in whole or in part by the CECR with funds from the U.S. Department of Education under contract number ED-06-CO-0110. The content does not necessarily reflect the position or policy of CECR or the Department of Education, nor does mention or visual representation of trade names, commercial products, or organizations imply endorsement by CECR or the federal government.

CECR

Center for
Educator Compensation
Reform

Allison Henderson, Director
Phone: 888-202-1513
E-mail: cecr@westat.com

34008.0312.83670507

D. Measurement.

What is inter-rater agreement and how can designers of teacher evaluation systems maximize it?

Inter-rater agreement is the degree to which two or more evaluators give the same rating to an identical observable situation (e.g., a lesson, a video, or a set of documents) using the same rating scale. Inter-rater agreement is often confused with inter-rater reliability. The latter concerns the similarity in the ordering of ratings made by two or more evaluators. While both apply to any situation where multiple observers assess an action or object (such as a teacher's performance), it is important to recognize their differences.

Agreement measures how frequently two or more evaluators assign the exact same rating: if both give a rating of "4," they are in agreement. Reliability measures the *relative* similarity between two or more sets of ratings, regardless of the *absolute* value of each evaluator's rating. Thus, even if two evaluators never assign the same numerical score, they could still have high inter-rater reliability if their ratings are in the same relative order (Tinsley & Weiss, 2000). The table below provides an example of how the two can differ.

	Low Agreement, High Reliability		High Agreement, High Reliability	
	Rater 1	Rater 2	Rater 3	Rater 4
Teacher A	1	2	1	1
Teacher B	2	3	2	2
Teacher C	3	4	3	3
Teacher D	4	5	4	4
Agreement	0.0		1.0	
Reliability	1.0		1.0	

In the example above, Raters 1 and 2 agree on the *relative* performance of the four teachers, but never on the *absolute* level of performance. In contrast, Raters 3 and 4 agree on both the *absolute* level and *relative* order of teacher performance. The inter-rater reliability between 1 and 2 is perfect, even though agreement is 0.

The same can apply to two raters scoring the same teacher. One rater's overall rating could be "2" and another's, "5," but if they agree on the teacher's *relative* strengths and weaknesses (for example, each gives a higher mark on classroom management than on content knowledge), reliability would still be high. To summarize,

agreement and reliability measure two separate and significant properties of an evaluation system: agreement, how often the ratings are the same, and reliability, the relative order of ratings.

Educator evaluation systems often designate certain rating levels as cutoffs for consequences or rewards, such as pay increases or retention decisions. In such situations, the educators being rated should probably be more concerned with inter-rater agreement than inter-rater reliability. Inter-rater reliability is still an important property, as a system with higher inter-rater reliability can more accurately determine a teacher's relative strengths and weaknesses. However, because reliability is concerned only with the relative order of rankings, it does not test for the possibility that some evaluators will be more severe than others. Since the high-stakes decision is generally made based on the absolute value of the rating (e.g., a "2" versus a "3"), inter-rater agreement is likely to be of greater concern to parties that may reap rewards or consequences based on their rating.

How can inter-rater agreement be maximized?

Because inter-rater agreement and inter-rater reliability both measure aspects of how consistently raters apply a rubric, each can be enhanced by reducing the variation between evaluators' ratings. Rater training is the most commonly identified method of decreasing the variability between raters, though studies have found that some variability can persist even after lengthy training (Hoyt & Kerns, 1999; Lumley & McNamara, 1995; Wang, 2010). Based on the finding that not all raters show reduced variability after training, some researchers have recommended recruiting more trainers than necessary and dismissing those who cannot pass agreement or reliability screenings (Henry, Grimm, & Pianta, 2010; Lumley & McNamara, 1995; Weigle, 1998).

Research has identified a number of components of effective training. When designers fail to allot sufficient time for training, it is much less likely to improve evaluator performance (Barrett, 2001). Training typically involves viewing artifacts or videotaped lessons that designers have scored, group discussion of the standards and their meaning, instruction on how to interpret language, information about common sources of systematic rater error, and practice scoring sample videos (Bakker, 2008; Borko & Stecher, 2005; Clare, 2000; Dymond, Renzaglia, Halle, Chadsey, & Bentz, 2008; Englehard & Myford, 2003; Henry et al., 2010; and others). Researchers often require that raters meet a predetermined minimum threshold of agreement or reliability in practice scoring sessions before they enter the field (Beesley, 2009; Dymond et al., 2008; La Paro, Pianta, & Stuhlman, 2004; Moon & Hughes, 2005; NICHHD ECCRN, 2005; and others). Based on their qualitative analysis of evaluators, Nijveldt, Beijaard, Brekelmans, Wubbels, and Verloop (2009) recommend that trainees discuss all of the relevant judgment processes, including personal biases and the mental processes needed to adjudicate between ratings or standards, so that raters will be more attuned to the ways in which their ratings might stray from objectivity.

Rater selection is also an important component of the process. The American Educational Research Association's standards advise that raters understand the domain they are assessing,

as well as the subjects to be assessed (American Educational Research Association, 1999). This suggests that raters may need to be drawn from a pool of people with subject matter expertise. Finding such raters can be challenging when evaluating high and some middle school teachers of world languages or advanced science. Rater attitudes and beliefs are also important. In their research on the Classroom Assessment Scoring System (CLASS), Henry et al. (2010) found that when raters' beliefs conflicted with the underlying theoretical foundation of the evaluation system, it was more difficult to bring their ratings in line with those of other raters. As with other training issues, one possible solution is to recruit more evaluators than needed so that those who are difficult to train can be excused.

Even if a good group of evaluators has been selected, maintaining inter-rater agreement and inter-rater reliability is not a one-shot deal, but an ongoing challenge. Rater agreement is not an inherent property of an evaluation system, but a facet that requires constant attention. Studies have shown that even reliable raters may change their rating behavior over time (Cogdon & McQueen, 2000; Harik, Clauser, Grabovsky, Nungester, Swanson, & Nandakumar, 2009; Myford & Wolfe, 2009). As a result of this finding, Lumley and McNamara (1995) recommend against the practice of certifying raters and then assuming that a single rater will produce consistent results—instead, they suggest periodic re-calibrations and the use of multiple raters.[1] The National Institute for Excellence in Teaching's evaluation system, The System for Teacher and Student Advancement, takes this approach (Daley & Kim, 2010).

[1] Lumley & McNamara focused on grading tests, not assessing teachers, but the same principle applies.

Holding raters accountable for accurate rating is another potential way to improve agreement. Researchers have found that raters, even experienced ones, provide much lower quality data when they do not know that their performance is being monitored (Romanczyk, Kent, Diament, & O'Leary, 1973; Weinrott & Jones, 1984). This could be accomplished by reviewing some of the raters' scores on artifacts or classroom observations, possibly by randomly double-scoring videotaped observations or artifacts. Dymond and colleagues (2008) found high levels of rater agreement between classroom observers and viewers of videotapes and contend that the use of videotaped observations could have the added benefit of reducing travel time and logistical issues, thereby making more observations possible.

Creating easily interpreted rubrics is another important element of improving inter-rater agreement and reliability. Capie (1978) found large increases in rater agreement when descriptors were included alongside rubric categories. Lumley (2002) showed that when evaluators are unsure how to adjudicate an issue based on the rubric, they resort to heavily weighing one aspect of the criteria or resorting to criteria that are not part of the rubric. This problem can be addressed through training, but also by refining the rubric in a way that addresses assessor difficulties in applying it. The National Board for Professional Teaching Standards (NBPTS) greatly improved its inter-rater agreement and rater reliability in the late 1990s in large part by providing more guidance to teachers about the rubric, as

well as overhauling the rubric by eliminating descriptors that had been identified as problematic (Wolfe & Gitomer, 2000).

Unlike the practices discussed above, rubric designers' choice of rating scale can present a tradeoff between agreement and reliability. Because of the nature of the measurements, rubrics that are designed to produce greater agreement may score worse on indices of reliability. For example, a "yes/no" checklist may produce high rates of agreement because ratings are often exactly the same, but such a rating system would produce lower reliability. This occurs because the limited choices in a dichotomous rubric leave little room to compare the relative rank order of different categories, which is how inter-rater reliability is determined (Tinsley & Weiss, 2000). By contrast, a rubric with more score points—say, 1–7 ratings instead of 1–4—will improve reliability by making relative comparisons easier, but reduces the likelihood of exact agreement on a particular score. Studies have found that wider rubric scales produce greater inter-rater reliability, but reduce rater agreement (Cook & Beckman, 2009; Penny, Johnson, & Gordon, 2000).

Perhaps the most critical part of designing an educator evaluation system that produces high reliability and agreement is a careful design approach that identifies problems and improves upon initial trials. According to Linn and Baker (1996), educational performance assessments are often designed and implemented before important methodological issues are considered. If sufficient time is given to develop a fair observation instrument, there will be fewer problems once it is implemented. The NBPTS was able to dramatically improve its reliability and agreement indices by modifying its training procedures to include more examples and more bias training, changing confusing components of its rubric, adding more guidance for teachers, and tweaking its rating scale (Wolfe & Gitomer, 2000). In addition to improving the agreement and reliability of the system, a lengthier design period allows stakeholders to iron out disagreements and build trust. Zellman and Perlman (2008) recommend devoting significant time to pilot programs, noting that post-implementation changes in observation systems led to "confusion and resentment" among parents and educators in several states. By making revisions based on pilot programs, designers can ensure the quality of an observation instrument before evaluators use it in the field, resulting in greater inter-rater agreement and smoother implementation.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Bakker, N. (2008). *Design and evaluation of video portfolios: Reliability, generalizability, and validity of an authentic performance assessment for teachers.* ICLON PhD dissertation series. Netherlands Organization for Scientific Research. Leiden: Mostert & Van Onderen.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, 2*(1), 49–58.

Beesley, A. (2009 April). *Measuring classroom assessment with a work sample.* Denver, CO: Mid-Continent Research for Education and Learning.

Borko, H., & Stecher, B. M. (2005 Dec). *Using classroom artifacts to measure instructional practices in middle school mathematics: A two-state field test.* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Capie, W. (1978 Nov). *The objectivity of classroom observation scales.* Paper presented at the Annual Meeting of the Southeastern Association for the Education of Teachers of Science (New Orleans, Louisiana, November 29, 1978).

Clare, L. (2000 Nov). *Using teachers' assignments as an indicator of classroom practice: CSE technical report.* Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing.

Cogdon, P. J., & McQueen, J. (2000 Summer). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163–178.

Cook, D. A., & Beckman, T. J. (2009). Does scale length matter? A comparison of nine- versus five-point rating scales for mini-CEX. *Advances in Health Sciences Education, 14,* 655–684.

Daley, G., & Kim, L. (2010). *A teacher evaluation system that works.* Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved August 29, 2011, from http://www.tapsystem.org/publications/wp_eval.pdf

Dymond, S. K., Renzaglia, A., Halle, J. W., Chadsey, J., & Bentz, J. L. (2008 Fall). An evaluation of videoconferencing as a supportive technology for practicum supervision. *Teacher Education and Special Education, 31*(4), 243–256.

Englehard, 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.* College Board Research Report No. 2003-1. New York: College Board Publications.

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009 Spring). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement, 46*(1), 43–58.

Henry, A. E., Grimm, K. J., & Pianta, R. C. (2010 Aug). *Rater calibration when observational assessment occurs at large-scale: Degree of calibration and characteristics of raters associated with calibration.* Dissertation. University of Virginia. Retrieved August 22, 2011, from http://proquest.umi.com/pqdlink?Ver=1&Exp=08-31-2016&FMT=7&DID=2396570251&RQT=309&attempt=1&cfc=1

Hoyt, W. T., & Kerns, M. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4,* 403–424.

La Paro, K. M., Pianta, R., & Stuhlman, M. (2004 May). The Classroom Assessment Scoring System: Findings from the prekindergarten year. *The Elementary School Journal, 104*(5), 409–426.

Linn, R. L., & Baker, E. L. (1996). Can performance-based student assessments be psychometrically sound? In J. B. Baron and D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities, Ninety-fifth Yearbook of the National Society for the Study of Education.* Chicago: University of Chicago Press.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19,* 246–277

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12,* 54–72.

Moon, T. R., & Hughes, K. R. (2005). Training and scoring issues involved in large-scale writing assessments. *Educational Measurement: Issues and Practice, 21*(2), 15–19.

Myford, C. M., & Wolfe, E. W. (2009 Winter). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale use. *Journal of Educational Measurement, 46*(4), 371–389.

National Institute of Child Health and Human Development, Early Child Care Research Network. (2005). A day in third grade: A large-scale study of classroom quality and teacher and student behavior. *The Elementary School Journal, 105*(3), 305–323.

Nijveldt, M., Beijaard, D., Brekelmans, M., Wubbels, T., & Verloop, N. (2009). Assessors' perceptions of their judgement processes: Successful strategies and threats underlying valid assessment of student teachers. *Studies in Educational Evaluation 35,* 29–36.

Penny, J., Johnson, R. L., & Gordon, B. (2000). Using rating augmentation to expand the scale of an analytic rubric. *The Journal of Experimental Education, 68*(3), 269–287.

Romanczyk, R. G., Kent, R. N., Diament, C., & O'Leary, K. D. (1973). Measuring the reliability of observational data: A reactive process. *Journal of Applied Behavior Analysis, 6*(1), 175–184.

Tinsley, H. E. A., & Weiss, D. J. (2000). Interrater reliability and agreement. In H.E.A. Tinsley & S.D. Brown, *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95–124). New York: Academic Press.

Wang, B. (2010 March). On rater agreement and rater training. *English Language Teaching, 3*(1), 108–113.

Weigle, S.C. (1998). Using FACETS to model rater training effects. *Language Testing, 15,* 263–287.

Weinrott, M. R., & Jones, R. R. (1984). Overt versus covert assessment of observer reliability. *Child Development, 55,* 1125–1137.

Wolfe, E. W., & Gitomer, D. H. (2000 January). *The influence of changes in assessment design on the psychometric quality of scores.* Princeton, NJ: Educational Testing Service.

Zellman, G. L., & Perlman, M. (2008). *Child-care quality rating and improvement systems in five pioneer states: Implementation issues and lessons learned.* Santa Monica, CA: RAND Corporation. Retrieved August 29, 2011, from *http://www.rand.org/content/dam/rand/ pubs/monographs/2008/RAND_MG795.pdf*