Research to Support the 5D+™ Rubric for Instructional Growth and Teacher Evaluation, Version 3* Annotated Bibliography

Appeldoorn, K. L. (2004). Developing and validating the collaboratives for excellence in teacher preparation (CETP) core evaluation classroom observation protocol (C.O.P.). (Doctoral dissertation). Retrieved from ProQuest (3154027).

In this dissertation, the researcher examined the internal consistency, reliability and validity of a classroom observation protocol (COP) in high school science classrooms. Observers were highly knowledgeable in science disciplines and participated in training using video observations. Reported reliabilities were 71.5% for intrarater reliability across multiple observations, and low interrater reliability ranging from 28-50% using Spearman's rho. The researcher attributed low reliability scores to inadequate training for observers and too much distance between training and actual observations. Recommendations were made for training and retraining observers, increased practice with real-life instruction before conducting actual classroom observations, and using cooperative groups where raters could discuss scoring procedures after rating independently.

Aspen Institute. (2011). Building teacher evaluation systems: Learning from leading efforts. Retrieved from website: http://www.aspeninstitute.org/sites/default/files/content/docs/education/AI_Perf%20Mgmt_Synthesis.pdf

This brief profiles two school systems that have recently overhauled their teacher evaluation systems – Washington D.C. Public Schools and Achievement First network of charter schools – in an effort to answer systemic-level questions for districts that are looking to improve teacher evaluation. Key findings include: the need to balance high-stakes accountability with ongoing support and feedback for teachers; the integral role of evaluator training; and the crucial need for multiple stakeholder ownership of performance management systems, with a special emphasis on teachers themselves.

Bell, C. A., Little, O. M., Croft, A. J., & Gitomer, D. H. (2009). *Measuring teaching practice: A conceptual review.* San Diego, CA: American Educational Research Association.

Researchers conducted an extensive literature review (reporting on results from 430 articles, published between 2002 and 2008) to assess validity evidence in observation protocols and other methods of evaluating teachers. They reviewed validity evidence for eight different observation protocols. Major findings on reliability include: studies were thin on documenting internal reliability; when they did, averages ranged from 0.6 to 0.8. Interrater reliability was reported unevenly; percentage agreement ranged from 0.81 to 0.96 on low-inference instruments, and "good" reliability for high-inference instruments was reported at 0.80. Double-coding procedures were rarely reported and documentation of rater bias was extremely thin. The review provides support for retraining of observers and conducting multiple observations, and argues that more study needs to be done on validity and reliability from sources external to the framework developers. The authors conclude that none of the observational protocols studied has strong enough validity evidence to be used one its own for high-stakes evaluation decisions.

---

*The research cited in the bibliography also supports the 5D Teacher Evaluation Rubric, Version 2.

Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J., Hamre, B., & Pianta, R. C. (2011). Within-day variability in the quality of classroom interactions during third and fifth grade: Implications for children's experiences and conducting classroom observations. ***The Elementary School Journal, 112*****(1), 16-37.**

This study, conducted by developers of the Classroom Assessment Scoring System (CLASS), sought to determine whether variability exists within a school day that would influence classroom observation results. They were specifically interested in environmental factors like time of day, activity setting, numbers of adults and students present, and instructional groupings. The authors concluded that within-day variability exists to some extent, especially as assessed by the Instructional Support component of the CLASS framework. They recommend scheduling classroom observations after the first twenty-five minutes of the school day and regulating environmental factors if possible, especially in time for transition. Regular documentation of observation schedules and environmental factors is also suggested to assist in understanding and refining consistency issues.

Curtis, R. (2011). ***District of Columbia public schools: Defining instructional expectations and aligning accountability and support.*** **Retrieved from Aspen Institute website: http://www.aspeninstitute.org/sites/default/files/content/docs/education/AI_DCPS_teacher%20evaluation.pdf**

This resource profiles Washington D.C. Public Schools' efforts to revamp teacher evaluation using Michelle Rhee's IMPACT evaluation system. Curtis tells the story of the system's development and implementation, and offers five major lessons for districts adopting new systems. D.C. Public Schools drafted and adopted its own instructional framework after researching and comparing existing frameworks. The report also details the train-the-trainer model that the district piloted, emphasizing analysis and calibration on video examples of instruction, multiple observations, and multiple observers to increase reliability. Interrater reliability results showed that given extensive training, principals and master teachers were rating with high (80-90%) reliability on most standards (as measured by average scores on a 1-4 scale). However, many teachers saw scoring as subjective and had negative perceptions of the evaluations' reliability.

Curtis, R., & Wiener, R. (2012). ***Means to an end: A guide to developing teacher evaluation systems that support growth and development.*** **Retrieved from Aspen Institute website: http://www.aspeninstitute.org/publications/means-end-guide-developing-teacher-evaluation-systems-support-growth-development**

This guide is intended for school systems in the process of designing or implementing reformed teacher evaluation systems, offering a step-by-step process informed by lessons learned from other systems. Step 5 ("identify capacity requirements") and Step 6 ("establish supervisor and system-level accountability for teacher growth and development") are especially relevant to this review. The researchers argue that evaluators must develop a daunting list of skills to achieve reliable and valid outcomes, and that districts should consider using peer evaluators, especially when considering the availability of time allocation for training and establishing reliability.

Daley, G., & Kim, L. (2010). ***A teacher evaluation system that works*** **(Working paper). Retrieved from National Institute for Excellence in Teaching website: http://www.tapsystem.org/publications/wp_eval.pdf**

This paper examines the structure of The System for Teacher and Student Advancement (TAP), a teacher evaluation system that combines classroom observations rated using TAP rubrics and value-added data to derive a cumulative score for teachers. These researchers make positive conclusions about the

effectiveness of the TAP system in positively and significantly correlating evaluation scores with student achievement growth and in leading to teacher growth over the study's two-year period. The paper is especially useful in specifically outlining procedures to establish reliability in classroom observations, but it does not report on precise reliability results. The study authors attribute the effectiveness of the TAP system to multiple observers, four to six classroom observations per year, extensive training including rater certification, and annual recertification. Criteria for certification are also outlined: evaluators must score within one point (on a scale from 1 to 5) on each of 19 indicators and within no more than two points on three indicators, as compared to expert ratings on videos. Leadership teams identify areas of discrepant scoring and work closely with evaluators as needed, working frequently on issues of interrater reliability.

Danielson, C. (2011). Evaluations that help teachers learn. *Educational Leadership, 68*(4), 35-39.

In this article, Danielson, author of the Framework for Teaching (FFT) classroom observation framework, highlights lack of credibility as a major problem inherent in traditional teacher evaluation systems. She discusses ways in which new evaluation systems are working to remedy issues of credibility and reliability, including training skilled evaluators, calibrating evaluator judgments, and conducting professional conversations about teaching and learning using the framework. Danielson cites Chicago Public Schools as a district that is actively working to merge the purposes of teacher evaluation and to develop a credible, accurate evaluation system. The article is a call to go beyond quality assurance toward sustainable teacher growth. It does not include specifics about reliability requirements.

Donaldson, M. L. (2009). *So long, Lake Wobegon? Using teacher evaluation to raise teacher quality.* Retrieved from Center for American Progress website: www.americanprogress.org/issues/2009/06/pdf/teacher_evaluation.pdf

Donaldson explores how school systems can best implement a valid and reliable teacher evaluation system, using Cincinnati Public Schools as an example of how to systematically overhaul teacher evaluation in a way that is meaningful and sustainable and more than just another initiative for teachers. Major features of Cincinnati's system are multiple observations, multiple evaluators including peers, ongoing extensive calibration work, and evaluator accountability. Donaldson offers seven recommendations to districts and states that are looking to reform teacher evaluation. These include robust professional development for evaluators, a blend of accountability and support for evaluators, and use of multiple evaluators to ensure checks and balances on the system.

Donaldson, M. L., & Peske, H. (2010). Supporting effective teaching through teacher evaluation: A study of teacher evaluation in five charter schools. Retrieved from Center for American Progress website: http://www.americanprogress.org/wp-content/uploads/issues/2010/03/pdf/teacher_evaluation.pdf

In this paper, Donaldson and Peske report on teacher evaluation practices enacted in five high-performing charter school systems across the country, in order to determine whether successful charter schools' relative freedom enables more effective teacher evaluation systems than do traditional school settings. Their findings reveal that charter schools are more focused on continuous improvement of teacher performance than summative assessment of individual teachers; thus, evaluation is positioned more as a professional habit than an administrative task. The researchers also highlight structural features that distinguish the evaluation systems of these five charter schools, including substantial training in classroom observation and providing feedback, and multiple evaluators and observations for

each teacher. Still, the researchers cite rater calibration as a challenge, especially given the time investment required.

Fry, R., & Ramsdell, R. (2011). *Enhancing teacher evaluation systems through effective classroom observation and tripod student surveys.* [PowerPoint slides]. Retrieved from Connecticut Association of Public School Superintendents website: http://www.capss.org/page.cfm?p=439

This presentation was given at the Teacher Evaluation that Works conference in December 2011. The authors cite examples of training schedules and lessons learned from six Florida school districts that are receiving Gates Foundation grants to pilot new teacher evaluation systems. Although the authors use the presentation to argue for extensive training and calibration work, they don't provide specific requirements for reliability or reports of reliability achieved using their methods. They also introduce their upcoming Calibration Engine, which the Gates Foundation hopes to make available for school districts and organizations wishing to analyze and calibrate evaluators and classroom observation data.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://www.tqsource.org/publications/EvaluatingTeachEffectiveness.pdf

This research synthesis of 120 studies examines how teacher effectiveness is currently measured, including features like validity and reliability, policy implications, and relevance of measures for formative and summative evaluation. The authors reviewed studies where the Framework for Teaching (FFT) was used, finding variation regarding the degree to which developer recommendations were followed and reliability for observers was established. They also concluded that the Classroom Assessment Scoring System (CLASS) has significant reliability evidence from elementary school work, but that the newer secondary tool does not. The researchers make several recommendations based on the studies reviewed, including the use of a validated observation protocol, extensive training for administrators, pre-establishment of rater reliability, and periodic recalibration to ensure consistency. They also argue that observations should occur several times per year in a combination of announced and unannounced visits, and that school districts should consider using a combination of administrators and peer evaluators, especially when specific content knowledge is required.

Goldstein, J. (2007). Easy to dance to: Solving the problems of teacher evaluation with peer assistance and review. *American Journal of Education, 113*(3), 479–508.

The author reports on findings from a four-year case study of a California school district that adopted Peer Assistance and Review (PAR), involving multiple stakeholders, including the local union. PAR provides an alternative way of supporting and evaluating beginning and struggling veteran teachers, with built-in accountability at all levels of the system, including oversight panels for the evaluators (Consulting Teachers, or CTs). After reviewing the program's implementation, Goldstein positions PAR alongside traditional principal-led teacher evaluation and argues that PAR is better situated to address structural barriers that typically inhibit successful and fair evaluation practices. Specifically, she cites issues like time, content expertise matching, rater calibration, collaborative work, and accountability as structural components that contribute to increased support and increased accountability for teachers.

Goldstein, J., & Noguera, P. A. (2006). A thoughtful approach to teacher evaluation. Educational Leadership, *63*(6), 31-37.

This article provides a brief overview of peer assistance and review (PAR) programs, which put peer evaluators in the roles of both providing individualized support and professional development for beginning and struggling teachers and serving as formal evaluators who can recommend retention or dismissal. The authors argue that districts that effectively implement PAR are more likely to see reliable evaluation results, due in part to strong accountability measures at all levels of the system.

Hamre, B. K., Goffin, S. G., & Kraft-Sayre, M. (2009). *Classroom assessment scoring system (CLASS) implementation guide.* Retrieved from Teachstone website: http://www.teachstone.org/about-the-class/

This document is particularly useful in detailing the training and calibration procedures for potential Classroom Assessment Scoring System (CLASS) evaluators. Key features of the way the framework developers strive to establish reliability include: standardized training for observers, a requirement for passing a reliability test within one point of master coders on 80% of all codes given using video classroom observations, and periodic recalibration requirements. The researchers also recommend that 5 to 15% of observations be double-scored to ensure a reliable evaluation system (they cite the example of a Virginia school system that has achieved 90% agreement within one point on the seven-point scale). They also emphasize longer observation times (up to two hours), standardizing time of day, and documenting and releasing reliability information to stakeholders to increase perceived credibility of the evaluation system.

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). Building a science of classrooms: Application of the CLASS framework in over 4,000 U.S. early childhood and elementary classrooms. Retrieved from Foundation for Child Development website: http://fcd-us.org/sites/default/files/BuildingAScienceOfClassroomsPiantaHamre.pdf

This paper presents research on the Classroom Assessment Scoring System (CLASS) and empirically tests this model across 4,000 classrooms. The researchers tested the degree to which the CLASS framework's three domains – emotional support, classroom organization, and instructional support – were consistent with observational data collected in actual classrooms. Reporting reliability evidence was part of this study. In this case, reliability was reported with Chronbach's alphas for internal consistency, falling into the "generally acceptable" range: alpha = .77-.89.

Henry, A. E. (2010). A*dvantages to and challenges of using ratings of observed teacher-child interactions. (*Doctoral dissertation). Retrieved from ProQuest (3462227).

This study examined the extent of rater calibration across more than 2,000 raters in a train-the-trainers model for Classroom Assessment Scoring System (CLASS). Henry found that it was possible to train large numbers of raters to calibrate to an observation tool, and that rater beliefs about teachers and children predicted the degree of calibration. 71% of potential raters passed a calibration test after the first round of training; raters whose beliefs aligned with the underpinnings of the CLASS system were more calibrated than those whose beliefs differed. Beliefs were found to be more important for calibration than levels of education or experience. The dissertation also provides detailed information about extensive training procedures for CLASS scorers, including interrater reliability requirements at 80% before certification is granted.

Humphrey, D. C., Koppich, J. E., Bland, J. A., & Bosetti, K. R. (2011). ***Peer review: Getting serious about teacher support and evaluation.*** **Menlo Park, CA: SRI International. Retrieved from Connecticut Association of Public School Superintendents website: http://www.capss.org/uploaded/Hard_Copy_Documents/Teacher_Evaluation_That_Works/PAR_PeerReviewReport_2011.pdf**

In this yearlong case study, the researchers profiled two California districts with well-regarded peer assistance and review (PAR) programs. Among other goals, they wanted to determine to what extent effective PAR programs support and measure teaching effectiveness, and to compare peer review programs with traditional teacher evaluation systems. The authors concluded that some of the unique features of PAR make it a highly viable alternative to traditional evaluation systems, including: the involvement of all stakeholders; an intensive blend of support and evaluation; frequent, targeted observations; and accountability at all levels of the system, particularly the use of governance boards to oversee the work of peer evaluators. The report provides an especially interesting section on the partnership between teachers' unions and the district, centered around peer review programs.

Marzano, R. G. (2011). The Marzano teacher evaluation model: Overview and resources. [PowerPoint slides]. Retrieved from Connecticut Association of Public School Superintendents website: http://www.capss.org/page.cfm?p=439

This PowerPoint presentation was used at the Teacher Evaluation that Works conference in December, 2011. It provides an overview of the Marzano instructional framework and classroom observation tool and introduces Marzano's approach to evaluating teachers through observations. The presentation cites the need for training and practice using the system, and offers information about how Florida school districts are compiling multiple measures to determine a cumulative score for teachers. This presentation, as with all of Marzano's available online materials, does not include many specifics about reliability requirements or calibration exercises.

Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at-scale." ***Educational Assessment, 13*****(4), 267-300.**

This study sought to determine the optimal number of observations and assignments needed to yield a reliable estimate of a teacher's practice in reading comprehension and mathematics, using the Instructional Quality Assessment (IQA). Generalizability and decision studies were conducted to determine the number of observations needed. For both content areas, as few as two observations yielded a reliable estimate of quality when teachers complied with the research requirements. Teachers were trained and were then observed on two consecutive days in the same class period by a single rater. Interrater agreement was assessed after training and prior to the study. The overall exact scale-point agreement between raters was 86% in reading comprehension and 82% in mathematics. The researchers recommend that potential observers undergo at least four days of training, demonstrate competency prior to rating teachers for real, and have adequate subject matter knowledge.

McCaslin, M., Good, T. L., Nichols, S., Zhang, J., Wiley, C. R. H., Bozack, A. R., … Cuizon-Garcia, R. (2006). Comprehensive school reform: An observational study of teaching in grades 3 through 5. ***The Elementary School Journal, 106*****(4), 313-331.**

In an observational study comparing two types of classroom approaches, the authors specifically attended to reliability issues. Their study has become a foundational work for other researchers looking

to establish adequate reliability. Using a leveled classroom observation tool, classroom observers participated in video and real-time training to attempt to calibrate scores with a minimum of 90% exact rater agreement using videos, and to achieve at least 80% interrater reliability when conducting actual observations. For 16 of 19 domains on the framework, exact rater agreement averaged 88%. Lower agreement on the other domains was attributed to confusing distinctions between levels on the framework or confounding observables, but even on those domains, the researchers achieved a range of 65-74% interrater reliability. The study has interesting implications for rater agreement using more complex frameworks like the Framework for Teaching (FFT), especially in consideration of the difficulty of achieving high levels of exact rater agreement.

Measures of Effective Teaching (MET) Project (2012). ***Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.*** **Seattle, WA: Bill & Melinda Gates Foundation (2012). Retrieved from website: http://www.metproject.org/reports.php**

This paper is the second in a series of updates about the Measures of Effective Teaching (MET) project launched by the Bill & Melinda Gates Foundation in 2009. The project's goal is to test new approaches to measuring effective teaching using a variety of available frameworks, in order to help school systems build fair and reliable systems for measuring teacher effectiveness. For this report, researchers examined five classroom observation instruments and calculated reliabilities using video libraries of instruction spanning thousands of hours of teaching. All potential raters completed 17 to 25 hours of training and had to pass a calibration exercise before beginning each observation day. The findings showed that scores from multiple raters and multiple lessons needed to be combined to achieve high levels of reliability. For four out of five observation instruments, researchers achieved 0.65 reliabilities by scoring four different lessons, each by a different observer. The authors made make strong recommendations that multiple observations and impartial observers should be used whenever high-stakes decisions are being made, and that observers should be required to demonstrate rating accuracy before participating in real observations.

Milanowski, A. T., Heneman, H. G., III, & Kimball, S. M. (2011). ***Teaching assessment for teacher human capital management: Learning from the current state of the art*** **(WCER Working Paper No. 2011-2). Retrieved from Wisconsin Center for Education Research website: http://www.wcer.wisc.edu/publications/workingPapers/Working_Paper_No_2011_02.pdf**

This paper reports on eight different classroom assessment systems that include observational components, in an attempt to delineate features that are recommended for designing and implementing state-of-the-art performance assessment for teachers. It lays out a strong rationale for establishing adequate reliability and compares the ways in which various assessment frameworks attempt to do this. The researchers conclude that almost all the approaches pay close attention to assessor reliability by requiring assessor training and accountability, including multi-day training, practice and certification requirements. They highlight catalysts for increased reliability, including using multiple evaluators, using evaluators with content-specific knowledge, scoring away from classroom distractions, using external evaluators and conducting multiple observations. They argue that these features are especially critical when high-stakes evaluation decisions are involved.

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation, 12*(1), 53-74.

This article distinguishes among varying types of reliability issues in classroom observation research. Muijs advocates for using multilevel models for statistical analysis, arguing that these methods are extremely helpful for detecting issues of validity and reliability. He cautions against researcher use of causal assumptions when highly controlled experimental designs are not used, which is the case in most school observation research. He also encourages researchers to use current statistical techniques for designing studies with adequate reliability. In light of classroom observation, Muijs advocates for extensive observer training and practice before formal evaluations are established.

National Council on Teacher Quality (2011). *State of the states: Trends and early lessons on teacher evaluation and effectiveness policies.* Washington, DC: National Council on Teacher Quality. Retrieved from website: http://www.nctq.org/p/publications/docs/nctq_stateOfTheStates.pdf

In this report, the National Council on Teacher Quality provides a glimpse into current state legislation, practices and pilots on teacher evaluation, including early observations on the implementation of new teacher evaluations in response to federal legislation and incentives. The report advocates that states establish system checks to ensure that teachers are being fairly evaluated. It also argues for trained third party evaluators whenever possible. The report is useful for comparing state initiatives, adopted frameworks, and training and evaluation practices.

Raudenbush, S. W., Martinez, A., Bloom, H., Zhuc, P., & Lina, F. (2008). *An eight-step paradigm for studying the reliability of group-level measures.* Retrieved from William T. Grant Foundation website: http://www.wtgrantfdn.org/publications_and_reports/browse_reports/raudenbush_1

This paper describes a method for conducting reliability studies and for planning impact studies that are methodologically sound when observing in classroom settings. The researchers argue that low reliability weakens the statistical power of evaluation. They detail several types of possible variance, including classroom variance, rater variance, rater-by-classroom variance, and segment variance. Their method for designing a reliability study addresses each type of variance. They caution that good reliability studies cross-classify sources like classrooms and raters in order to estimate the variation attributable to each and to their interactions, rather than nesting raters in classrooms. Interpreting the interactions between variables can help study designers to determine the effects of various techniques intended to increase reliability, like adding more raters, increasing observation frequency or duration, and increasing sample size.

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation. Retrieved from University of Chicago Consortium on Chicago School Research website: http://ccsr.uchicago.edu/publications/rethinking-teacher-evaluation-chicago-lessons-learned-classroom-observations-principal

This document is an extensive report on the first year of implementation of a new teacher evaluation system in Chicago Public Schools. It is a powerful example of documentation, providing a coherent narrative and specific analysis of implementation and validity results, and making recommendations for refining the system. The study offers specific reliability analysis, concluding that the classroom observation ratings were reliable measures of teaching practice, with a few important caveats. Principals

were generally more lenient raters than external observers. Reliability was highest at the low end of the rating scale. The report concludes with a design guide for school districts piloting their own teacher evaluation systems, with special emphasis on the logistics of observation procedures, training, and evaluator feedback and accountability.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., & Benford, R. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics, 102*(6), 1–20.

This study was conducted using the Reformed Teaching Observation Protocol (RTOP) to assess teaching quality in secondary and college-level mathematics and science classrooms. It is notable in that over a two-year period of data collection using this high-inference instrument, the research team was able to achieve high levels of interrater reliability (0.95) and internal consistency as estimated by Cronbach's alpha (0.88 to 0.97). Researchers attributed high reliabilities to the way that training was conducted, including features like cycles of observation using video examples, discussion and inquiry among team members, and double scoring of observations.

Sterbinsky, A., & Ross, S. M. (2003). *School observation measure reliability study.* Memphis, TN: The University of Memphis, Center for Research in Educational Policy.

After reviewing available reliability studies for classroom observation instruments, the authors concluded that some evidence of reliability could be found, but only in specific contexts. This study attempts to consider contextual variables in other reliability studies by using previous conclusions about number of observers, frequency of observation, environmental factors, types of observers, and characteristics of classroom observation tools. Using lessons learned from previous research, Sterbinsky and Ross created a new reliability study on the School Observation Measure, a relatively low-inference classroom observation protocol. They provided specific training for observers and calculated both internal reliability and interrater reliability using a generalizability study (G study) and a decision study (D study). Results indicated that the number of observations was a significant factor that changed reliability: from a phi coefficient of 0.38 for one observation to 0.82 for eight observations. The authors recommended a minimum of five school observation measures (SOMs) at the elementary level and a minimum of eight at the secondary level to achieve adequate reliability.

Stuhlman, M. W., Hamre, B. K., Downer, J. T., & Pianta, R. W. (2010). A practitioner's guide to conducting classroom observations: What the research tells us about choosing and using observational systems to assess and improve teacher effectiveness. Charlottesville, VA: The Center for Advanced Study of Teaching and Learning, University of Virginia.

This multi-part research synthesis offers several key cautions for classroom observation frameworks and their use within teacher evaluation systems, especially for high-stakes decision making. The researchers argue that the instruments used must be subjected to extensive testing and evaluation and that this process is just beginning. They emphasize two important aspects of reliability – stability over time and consistency between observers – and make recommendations for achieving each aspect. A major conclusion of this report is that reliable observations of classroom instruction require a substantial time commitment, for both the training of evaluators and for the observations themselves. The researchers also highlight many specific recommendations for establishing adequate reliability, including using multiple raters, minimizing bias, enacting system and evaluator checks, conducting multiple observations,

and documenting the way observations are done so that environmental factors can be standardized to the extent possible.

Stumbo, C., & McWalters, P. (2011). Measuring effectiveness: What will it take? *Educational Leadership, 68*(4), 10-15.

The authors describe seven challenges that school systems face in implementing new teacher evaluation systems. One key challenge involves the training of evaluators, which the authors argue is a crucial factor in determining whether evaluations will be credible and reliable. Stumbo and McWalters highlight inter- and intrarater reliability as important features of evaluation systems, but caution that standards for reliability have not yet been regularly established. They suggest increased partnership between research and assessment experts, policymakers and instructional leaders to implement good systems, and they advocate analyzing lessons from the business field to develop team-based accountability systems that better reflect the realities of professional collaboration.

The New Teacher Project (2011). *Smart spending for better teacher evaluation systems.* Brooklyn, NY: The New Teacher Project. Retrieved from website: http://tntp.org/assets/documents/TNTP_Smart_Spending_2011.pdf

This report reviews teacher evaluation roadmaps for states that have received Race to the Top funding and makes recommendations for moving beyond evaluation design into implementation, with a focus on five key areas: tools and systems, training, communications, monitoring and support, and sustainability. The report cautions that training evaluators for fair and consistent results will require a significant investment of both one-time and ongoing resources. The report also points to the need for system monitoring, evaluator accountability, and detailed documentation, so that changes can be made to refine evaluation systems toward valid, reliable outcomes.

Tyler, J. T. (2011). *Designing high quality evaluation systems for high school teachers: Challenges and potential solutions.* Retrieved from Center for American Progress website: http://www.americanprogress.org/wp-content/uploads/issues/2011/11/pdf/high_school_evaluation.pdf

Tyler's report advocates for a focus on teacher evaluation effectiveness at the high school level. He argues that implementing reliable and valid classroom observation systems will be challenging in light of the difficulty of matching observers and content areas, which may decrease both reliability and teacher buy-in. In his recommendations for classroom observations at the high school level, Tyler cites Cincinnati's use of peer evaluators who have content and pedagogical knowledge and can help make fair and accurate observations. Tyler argues that this use of human capital is paying off in terms of tangible increases in teacher effectiveness as a result of well-designed practice-based evaluations. He also views video instructional examples as effective for cost savings, increased reliability, and national pairing of teachers and content-area specialists who could serve as evaluators.