Annotated Research Bibliography

This annotated bibliography summarizes key research literature on the following four topics: (1) rater agreement; (2) goal-setting linked to student outcomes or test scores; (3) using frameworks, rubrics, or common language to improve student learning; and (4) implementation of evaluation systems. It is intended to facilitate policy decision-making in Washington State on teacher and principal evaluation.

1. Rater Agreement


This Aspen Institute paper discusses the design, training, and implementation of Washington DC's teacher evaluation system, IMPACT. In their teacher evaluation system, teachers were observed five times throughout the year for 30 minutes per observation. Principals conducted three of the observations, and Master Educators conducted two of the observations, with each conducting one announced observation. For training, the Master Educators participated in a six-week training orientation on Washington DC's Teaching and Leading Framework, which guides the classroom observations. During this training, observers participated in learning how to build their skills in observing classroom instruction, and how to provide written and verbal feedback using their observation notes. During this training they spent hundreds of hours reviewing videos of classroom instruction and norming their ratings. Along with their training, Master Educators met twice a month for two hours in order to maintain calibration, problem-solve, and discuss areas of concern. Results determined that out of the 13 standards that were used for classroom observations, principals and Master Educators level of variation was quite small (0.2 or less) on 10 of the standards. Thus, Master Educators and principals agreed on the scores that they were providing their teachers. In order to improve on the other standards, principals and Master Educators participated in additional training to reduce the level of variation and calibrate their scores on the other standards. However, it should be noted that even though there are similar levels of variation, the principal and Master Educator’s scores only correlated at 0.57, which is a moderate correlation.


The goal of this paper by Charlotte Danielson is to describe four steps in training principals on how to become effective evaluators and obtain a certain level of inter-rater reliability. (1) Principals need to become very familiar with the observation instrument. (2) Principals need to learn how to recognize sources of evidence. What does practice look like across the different domains? (3) Principals need to learn how to interpret the evidence collected against the rubric and performance levels. Even after training principals on classroom observation instruments, they still need multiple opportunities to continuously practice using the framework, and calibrate with colleagues. As such, (4) principals should calibrate scores they provide teachers with their colleagues, and discuss with colleagues how created similarities and differences in both scores, and sources of evidence collected.

The goal of the Measures of Effective Teaching (MET) study by the Bill & Melinda Gates Foundation was to understand various sources of measurement error and the resulting reliability scores for teacher observations, while comparing the scores of different types of observers. The study was conducted in Hillsborough County with administrators as well as school- and district-based peers observing 67 teachers who provided four different lessons. Results demonstrate that administrators and peers were similar in their scores in many ways, but also had some differences. First, there was not a great deal of variation among scores across observers. Using a four-point scale, observers rarely scored teachers in the bottom or top categories (5 percent and 2 percent, respectively). Even though administrators provided more differentiated scores than peer observers, administrator’s rankings of teachers inside their school were similar to the rankings produced by observers outside their building (0.87 correlation, adjusting for measurement error). Even with these similarities, the low distribution of scores made the small differences in administrator rankings have large effects (i.e., a 0.1 point difference in scores could move a teacher by 10 percentile points in rank).

Based on their results, the authors provide recommendations to improve the reliability of observations. First, multiple observations increased reliability. However, the increase in reliability is about twice as large when a second observation is conducted by a second observer as compared to the same observer. This is because it reduces two sources of error, lesson-to-lesson variance, and rater-by-teacher error variance. Only lesson-to-lesson variance is reduced when a single observer is used. Additionally, reliability improved when teachers were able to choose their own video that was observed. Finally, it should be noted that rater effects are higher in this study compared to other MET studies. In other MET studies, raters participated in daily calibration exercises before scoring, a practice that was not required in this study.


This paper finds that in order to train principals and evaluators programs should have participants: (1) define and understand the difference between bias, interpretation, and evidence; (2) apply the knowledge and skills they learned in training to the grades and content areas they will be assessing; (3) use exemplar videos that have been pre-scored by experts; (4) score authentic lessons alongside an expert; and (5) pass certification tests that require participants to score accurately and identify appropriate reasons for a score. Certification tests should be required of evaluators, as they demonstrate evidence of the evaluator’s ability to use the rubric in an “accurate, fair, impartial, and consistent” manner.

In addition, the paper notes that along with initial training, evaluators need to (1) review the rubric frequently; (2) be reviewed on their application of the rubric; and (3) receive feedback on their application of the rubric and any potential misunderstandings. Although teachers do not need to be trained as extensively as evaluators, it is suggested that they participate in the core training, as teachers are more likely to buy into the system if they know what they are being evaluated on.

---

1 The authors define reliability as “proportion of variance in observed scores that reflects persistent differences between teachers” (p. 18).

This rigorous study from the Consortium on Chicago School Research describes the pilot of the teacher evaluation system in Chicago Public Schools, which included 101 schools, 501 teachers, in the 2009–10 school year. The pilot used Danielson’s Frameworks for Teaching, in which the principals conducted two formal classroom observations, with pre- and post-observation conferences. They found the observation to be a reliable tool to assess teacher performance, in that principals and trained observers provided teachers the same ratings. This is particularly true at the low and middle range, as principals tended to give more “Distinguished” ratings than did trained observers.

In qualitative interviews, principals admitted that they rated some teachers as “Distinguished” even though they should have been rated as “Proficient” in order to maintain relationships. They felt that it was better to coach these teachers rather than try to explain the difference between “Proficient” and “Distinguished.” However, when correlating principal’s observation scores with teacher value-added scores, a teacher rated as Distinguished by the principal did have higher-valued added scores. Thus, it is not necessarily true that the observer was providing the correct score.

2. Goal-Setting Linked to Student Outcome or Test Scores


This study focuses on the effects of a learning intervention—“Cover-Copy-Compare”—with and without the preestablishment of objectives. The study compared the effects of the intervention alone, the combination of it with goal setting, to a control condition on the mathematics fluency of 173 third-grade students. Treatment sessions occurred twice weekly for a total of six weeks, and multilevel modeling was used to examine progress across intervention sessions. Results of the analysis showed that combining Cover-Copy-Compare with goal setting leads to steeper slopes and higher final scores than most other treatment conditions (higher mean score among goal-setting group (36.14) over the mere intervention group (31.10)). In addition, assessment of student retention was taken one week after the intervention and again one month after the intervention. The effects were retained for one month and generalized to similar stimulus conditions. This finding suggested a long-term impact of the goal setting on student academic performance.


This paper describes a quasi-experimental study of 151 sixth-grade students, who worked with a mathematics learning software program during regular classroom instruction. Students were divided into three conditions: control group; a feedback-only group where students received attributional software-generated feedback; and a feedback plus metacognitive goal-setting group where students received the attributional feedback and additional self-regulation training designed to foster metacognitive control. The results of the quasi-experimental study indicated that the metacognitive group did outperform other
groups in the one-week assessment of mathematics skills. The promotional effects on the components of student motivation associated with attributional feedback could be detected immediately after the training period and were largely sustained for five months. However, when considered for the entire school year (i.e., analyzed the follow-up data) no promotional effect of attributional feedback on mathematical knowledge could be proven statistically. This result implied the immediately positive effects of metacognitive goal setting with no long-term improvements.


The report conducted a meta-analysis on setting objectiveness in the instructional practices. The author searched literature from both educational and psychological databases. Four studies (Codding et al., 2009; Dresel & Haugwitz, 2008; Glaser & Brunstein, 2007; Perels et al., 2009) were included in the meta-analysis. All studies produced positive effects on objective setting with an overall effect of $g = 0.31$. The study reported the form of objective setting included the establishment of goals, metacognitive skills, and self-regulation. In addition, there was mixed evidence that setting objectives had lasting effects on material retention. In general, the findings supported the hypothesis that setting objectives is a robust instructional strategy for improving student learning. Furthermore, the study recommended the strategy of supporting students in self-selecting learning targets, self-monitoring progress, and self-assessment.


This report reviewed three studies (Wise & Okey, 1983; Walberg, 1999; Lipsey & Wilson, 1993) on the issue of goal setting. Studies synthesized showed percentile gains in student achievement ranging from 18 to 41. In addition, the report drew the following three conclusions from the research on goal-setting:
(1) Instructional goals narrow what students focus on. Therefore, although students generally score higher on the instruction related to the specific academic goal, they likely would score lower (about 8 percentage points) on information that is incidental to the goal, but still covered in the class.
(2) Instructional goals should not be too specific. In other words, instructional goals stated in behavioral objective format do not produce student learning gains as high as instructional goals stated in more general formats. (3) Students should be encouraged to personalize the teacher’s goals. Once classroom academic goals are set, students should be encouraged to customize them to fit their personal needs.


This study used a quasi-experimental design with a hierarchical linear model to examine the relationship between goal setting and student achievement in the high school Spanish language classroom across time at both the individual student and teacher levels. The intervention—LinguiFolio—was a portfolio that focused on student self-assessment, goal setting, and a collection of evidence of language achievement. The intervention was introduced into 23 high schools with a total of 1,273 students. The LinguaFolio goal-setting process involves students directly in the learning process as they keep track of learning goals and track progress toward these goals. Students identify goals, create an action plan that details the path to goal attainment, and provide examples of their work that serves as evidence of accomplishment of learning goals. A final step involves student reflection to determine at what level the goal was
accomplished. The results of the correlational analysis of the goal-setting process and language proficiency scores were the student’s 0.17 increase of the writing score at teacher-level and 0.283 increase at student level. The results reveal a statistically significant relationship between the goal-setting process and language achievement (p < 0.01).

3. Using Frameworks, Rubrics, or Common Language to Improve Student Learning


This report presents results from a randomized controlled trial that used an online platform, My Teaching Partner-Secondary (MTP-S), as a professional development tool. MTP-S uses the Classroom Assessment Scoring System-Secondary (CLASS-S) to improve the student-teacher interactions in the classroom. The tool uses the CLASS-S domains in order to describe the current behaviors and interactions that occur in the classroom, as well as target points to intervene. The MTP-S interventions include an initial workshop, annotated video library and a personalized coach. During the academic year, teachers record teaching and upload it into the system. The personalized coach reviews the lesson, noting areas where there are positive interactions and areas in which the teacher could grow. In other words, teachers are provided notes at distinct moments within the video. Once teachers review the comments, they participate in a phone conference with their coach, using the CLASS-S as a focal point to improve student-teacher interactions. Teachers who participated in the intervention, My Teaching Partner, produced substantial gains in student achievement a year after the intervention occurred, moving students from the 50th to 59th percentile (there were no significant changes in student growth during the year of the intervention). These achievement gains were facilitated by improved student-teacher interactions, regardless of content area.


This paper from the Consortium on Chicago School Research focuses on a pilot of the teacher evaluation system in Chicago Public Schools, which included 101 schools, 501 teachers, in the 2009–10 school year. The pilot used Danielson’s Frameworks for Teaching, in which the principals conducted two formal classroom observations, with pre- and post- observation conferences. The new evaluation system produced many benefits. Many principals and teachers thought that the conferences were more objective and focused on actual classroom instruction. Principals and teachers felt as though there was a common language that could be used to talk about instruction. Teachers liked that they had something that they could work toward, and had concrete teacher practices that they could reflect on whether they were or were not meeting those targets. Teachers and principals felt that the pre-conference also helped with the planning process and helped them to be more reflective about how the teacher would approach the upcoming lesson.

Although there was general positive feedback, the positive attitudes toward the conferences depended on the principal’s ability to provide feedback, in which many principals lacked the necessary skills to provide targeted instructional coaching to their teachers. Even though there was some variability in the types of coaching, the majority of conferences were dominated by principal talk (75 percent of conference time was principal talk), in which the principal provided few higher order questions (approximately 10 percent of questions). Principals were also concerned that the conferences took too much time, which led
principals to not engage in the conferences and observations as much. Not only was providing feedback difficult for some principals, but identifying the evidence for scores was still difficult for some. Principals also varied how they asked questions depending on the needs of the teacher, or the ability of the teacher to be a reflective practitioner.


Evaluations have the potential to improve teacher performance in two distinct, but overlapping ways: receiving incentives for good performance, and the investment they bring to human capital. However, performance will only increase depending on how well the instrument teachers are being evaluated, assessing those teaching behaviors that produce significant gains in student achievement. This study focuses on the Teacher Evaluation System (TES) in Cincinnati Public Schools in the 2000–01, as it is one of the well-developed teacher evaluation systems. It evaluates teachers on multiple classroom observations (based on Danielson’s Frameworks for Teaching) and evidence of practice (e.g., lesson plans, professional development activities, and family contact logs), but not student test scores. In regards to observations, teachers are scored four times throughout the year by highly trained observers, three times by a peer evaluator, and once by their school principal. The study focused on the impact TES had on mid-career teachers’ performance. Results from the TES study demonstrate that mid-career teachers who participate in the evaluation have improved student achievement growth (as measured by and end-of-the-year state test) in the evaluation year and the year following the evaluation. This improvement in teaching is particularly evident for those teachers who were less skilled prior to the evaluation.

4. Implementation of Evaluation Systems


This brief from the Bill & Melinda Gates Foundation provides a summary of the findings of the MET study. More specifically, this brief answered three key questions related to measures of effective teaching, and how these measures can be effectively used in a teacher evaluation system.

The first question asked, “Can measures of effective teaching identify teachers who better help students learn?” Using random assignment of students to teachers in the 2010–11 school year, the MET study found that teachers who were identified as more effective in the 2009–10 school year (using a combined measure of teacher effectiveness) had students with higher achievement, both on standardized tests and on more challenging assessments in English and math.

The second question asked, “How much weight should be placed on each measure of effective teaching?” They found that student achievement should be between 33 to 50 percent of a teacher’s composite score to be the most stable from year to year and most predictive of multiple measures of student achievement (more challenging assessments and the standardized tests). Thus, they recommend that student achievement should never be used alone, but need to be combined with multiple other measures (in this case student perception surveys and classroom observations).

Finally, the third question asked, “How can teachers be assured trustworthy results from classroom observations?” Observation scores tend to vary more from observer to observer than they do from lesson to lesson; thus, more than one observer produces more reliable scores. For this third question, they had administrators and peer observers in one district provide scores for more than 3,000 lessons. They
learned that (1) having an additional observer observe shorter video segments (or parts of a lesson) is a cost-efficient way to increase reliability of teacher scores; (2) teacher’s own administrators discern bigger differences between teacher practice, which increases reliability (even though administrators tended to rate their teachers higher than outside observers); and (3) additional outside observers would help decrease in-school bias, but this would only have to be done with a subset of teachers.


This study focuses on a Peer Assistance and Review (PAR) program’s oversight panel, and how an oversight panel influences the process of teacher evaluation systems. In order to accomplish this goal, a case study of a district in California (more than 100 schools and 3,000 teachers) provides information about how peer coaches (who were released from their full-time positions to assist with teacher evaluations) interacted with the oversight panel to conduct teacher evaluations and make personnel decisions. These coaches, rather than being school based, reported to a districtwide oversight panel, and were tasked with both mentoring and summative evaluations (generally of newer teachers). The oversight panel generally consists of teacher union leads and director of human resources (or other high ranking district administrators). The oversight panel influences teacher evaluations in two ways. First, they provide guidance and coaching to the peer coaches in how to be an effective mentor and evaluator. Second, the peer coaches have to justify their ratings, both for why they believe that teachers should be re-hired or non-renewal.

This new type of teacher evaluation process has multiple types of benefits. First, it allowed for more transparency about the teacher evaluation system because the evaluators and people making the decisions were meeting regularly and having discussions about the teachers they were evaluating. Additionally, it created a professional learning community between teachers and district personnel focused on teacher practice. With that, there was more distributed responsibility across the evaluation, so that not only one person was responsible for the personnel decisions. Additionally, it created a sense of collective responsibility to uphold professional standards of practice. As such, the PAR program provides a good model on how to make teacher evaluations transparent and collaborative to improve teacher growth.


This paper from the National Education Association reviews five teacher evaluation systems that are believed to comprehensive and innovative in their approach to evaluating teachers. These programs were selected because of their comprehensive design; they were designed to improve instruction, support teachers, increase retention, and improve student achievement. After reviewing each of the five teacher evaluation systems, Little states three overarching implementation considerations: (1) the need to ensure that the system is credible to all stakeholders. This can be facilitated by using accepted standards of performance, valid instruments, training, and continuously calibrating raters, ensuring teachers are receiving appropriate feedback, targeted support, and that there is open communication between evaluator and teacher. It is also recommended to have opportunities for teachers to advance within their teaching career. (2) Leadership is a critical ingredient. Administrators need to buy into the system in order to implement the evaluation system with fidelity. As such, principals need to be held accountable as well. (3) Implementing evaluation systems cannot be rushed; appropriate time and resources need to be provided to districts in order to implement them effectively and refine them along the way.

---

2 The five systems include: (1) Teacher Advancement Program (TAP); (2) Framework for Teaching (FFT); (3) Professional Compensation System (ProCamp); (4) Peer Assistance and Review (PAR); and (5) Beginning Educator Support and Training Program (BEST).

Teacher evaluations have the potential to leverage instructional improvement, using teacher evaluations as a dual focus as a formative and summative assessment. Not only will teacher evaluations provide teachers feedback on their performance, it better assists targeting professional development for teachers. This Center on Great Teachers and Leaders brief provides multiple recommendations on implementing evaluations. First, teachers are more receptive to evaluators who have knowledge of the curriculum, content, and instruction they provide students, making peer evaluators a cost-effective solution to this recommendation. Second, both tenured and non-tenured teachers should receive frequent evaluations. Although there is not a general consensus about the number of observations, some scholars suggest four or five observations. Third, training evaluators is critical to obtain reliable and unbiased estimates. Finally, the teachers should be communicated with before, during, and after the evaluation process. The brief provides further examples and research into how these recommendations are used in practice and have been researched.