Best Practices in the Evaluation of Teaching

IDEA Paper #69 • June 2018

Stephen L. Benton, The IDEA Center • Suzanne Young, University of Wyoming

Abstract

Effective instructor evaluation is complex and requires the use of multiple measures—formal and informal, traditional and authentic—as part of a balanced evaluation system. The student voice, a critical element of that balanced system, is appropriately complemented by instructor self-assessment and the reasoned judgments of other relevant parties, such as peers and supervisors. Integrating all three elements allows instructors to take a mastery approach to formative evaluation, trying out new teaching strategies and remaining open to feedback that focuses on how they might improve. Such feedback is most useful when it occurs in an environment that fosters challenge, support, and growth. Rather than being demoralized by their performance rankings, faculty can concentrate on their individual efforts and compare current progress to past performance. They can then concentrate on developing better teaching methods and skills rather than fearing or resenting comparisons to others. The evaluation of teaching thus becomes a rewarding process, not a dreaded event.

Keywords: Evaluation of teaching, summative evaluation, formative evaluation, mastery orientation

"Evaluation without development is punitive,
and development without evaluation is guesswork."
—Michael Theall (2017, p. 91)

After making a presentation about faculty evaluation at a small college, a consultant receives an invitation to speak with the dean, who is concerned about an individual faculty member. In an inner office, the dean shows the consultant multiple class reports of student ratings across several years for the same anonymous instructor. The consultant studies the reports and surmises that the faculty member consistently scored below average on every course. The dean asks, "What should we do?" The consultant quizzically responds, "Well, what *have* you done?" The dean's telling response is, "Nothing, that I'm aware of."

Like the dean in this example, most administrators in higher education examine various sources of evidence regarding their instructors' teaching effectiveness. Those sources, which may include student ratings of instruction (SRI), instructor self-assessments, peer/supervisor review of course materials, and other measures (e.g., peer/supervisor classroom observations, alumni ratings), are often used to evaluate performance and to make decisions about merit-based salary increases, faculty retention, tenure, and promotion.

However, similar to the dean's response, confusion often reigns about how to best use evidence to help faculty improve. The fundamental purpose of evaluating teaching is, after all, to provide the best instruction for students (Bana e Costa & Oliveira, 2012; Braskamp & Ory, 1994; Lyde, Grieshaber, & Byrns, 2016; Joint Committee on Standards for Educational Evaluation [JCSEE], 2009; Svinicki & McKeachie, 2014). Whether making summative decisions about retention and promotion or helping faculty become better teachers, the ultimate goal of evaluation is the same—to improve student learning.

This paper is intended for faculty and administrators who want to learn how to apply principles of evaluation to both formative and summative decisions about teaching effectiveness. We begin by making an important distinction between assessment and evaluation. We then discuss criteria for judging the

trustworthiness of evidence, the difference between summative and formative evaluation, their respective benefits, and the value of adopting a growth mind-set for evaluating instructors. Finally, best practices in the evaluation of teaching are described.

Distinguishing Assessment from Evaluation

The distinction between assessment—which is the process of observing evidence of teaching behavior and drawing inferences about abilities—and evaluation, which involves judgments about how well one has taught and about how to improve, is an important one. In our view, SRI are but one source of evidence that should be considered in the overall evaluation of teaching. There are many aspects of college teaching that students are not qualified to judge (e.g., course content, instructors' subject-matter knowledge, course design, course goals, quality of student evaluation; Berk, 2018; Benton & Cashin, 2011). Overall evaluation is thus best left in the hands of those authorized and qualified to make judgments—faculty and administrators. Accordingly, a course evaluation should be an ongoing process, not an event confined to a single class period near the end of a semester.

In preparation for evaluating teaching, faculty should collect relevant evidence, both quantitative and qualitative. Although objectivity is important, the most crucial question is whether the evidence collected can be used to advance the professional development of the individual faculty member and improve instruction (Seldin, 2006; Svinicki & McKeachie, 2014). Before using any tool or method to assess teaching, then—be it SRI, peer/supervisor review of course materials, or instructor self-assessment—one should ask, "What is its intended purpose? How will the information help improve teaching or student learning?" As baffling as it may seem, educators often distrust tools that actually have substantial evidence of validity and reliability, such as well-designed SRI (Benton & Ryalls, 2016), whereas they sometimes place more trust in data that lack such evidence, such as students' written comments and faculty hearsay (Braskamp & Ory, 1994; Hativa, 2014).

Weigh the Trustworthiness of Evidence

To conduct evaluations without sufficient evidence is poor practice, but to do it without using credible measures is not much better. When deciding which sources of evidence to include, four conditions of trustworthiness should be considered—reliability, validity, fairness, and social consequences (Berk, 2006; Bradley & Bradley, 2010; Braskamp & Ory, 1994).

Reliability refers to the consistency in scores across repeated instances when using a particular measure (American Educational Research Association [AERA], 2014). Reliability means that if the same measure were administered at different times with the same group of people, the results would generally be the same. Related terms include *consistency, stability,* and *generalizability.*

Reliability is especially important when the consequences of the teaching evaluation will have a lasting impact. If decisions resulting from an evaluation cannot be easily reversed, such as those regarding continued employment, tenure, promotion, and merit pay, a high degree of reliability is necessary. If, on the other hand, an erroneous initial decision can be reversed (e.g., a change in a teaching method), a merely moderate level of reliability may be acceptable.

Several factors can affect reliability. First, evaluations based on multiple measures tend to be more reliable (on average) than those based on only one or a few (AERA et al., 2014). Consequently, well-constructed SRI instruments tend to have high reliability because they are based on the perceptions of multiple raters (students) across multiple occasions (class sessions). Second, rater preparedness can impact reliability. All raters (whether students or peer faculty), for example, should be given the same set of instructions for completing the evaluation. Otherwise, the differences in instructions introduce irrelevant variability into the responses. Third, variations in the physical environment can also influence the results. When collecting student ratings, therefore, it is best that all students fill out the form at the same time in the same format (e.g., all on a mobile device during class).

Whereas reliability concerns consistency, *validity* refers to whether interpretations made from the measure are appropriate and whether the measure is being used for its intended purpose (AERA et al., 2014). Evidence can be reliable but not necessarily valid. For instance, peer faculty observers might be highly consistent in their

observations about whether an instructor greets students at the beginning of class. However, what does greeting students necessarily demonstrate about effective teaching or about how much students have learned?

In practice, validity is tied to use. A tool is not valid per se but depends on how one interprets the score and uses it to make decisions. In other words, are the interpretations and uses justified by the accumulated evidence? For example, let's consider the functions of a hammer. Hammers vary in type and purpose. A ball-peen hammer is useful for metalworking; a framing hammer works best for framing wooden houses; a dead-blow hammer is used to minimize damage to a surface, such as tile. The validity, or effectiveness, of each hammer depends on its being used appropriately. The same principle applies to measures of teaching effectiveness. Student ratings provide useful feedback about what actually occurred in the classroom; peer review of course materials provide knowledgeable input specific to an academic discipline or type of pedagogy; instructor self-assessments ideally engender self-reflection and self-improvement. But just as no type of hammer is good for all situations, no single measure is adequate as evidence for evaluating teaching.

Whereas reliability and validity address the quality of the evidence, fairness protects the individual by ensuring that the measures collected adequately represent the complexity of teaching and its many outcomes. The best way to ensure fairness is to include multiple measures (e.g., SRI, instructor self-assessments, review by peers/supervisor), which prevents a single source (e.g., student ratings or a single peer observer) from skewing the evaluation. Too often faculty justifiably lose trust in a system that relies too heavily on a single measure, such as student ratings (Berk, 2018).

Another aspect of trustworthiness is the measure's social consequences, or its intended and unintended effects. When instructors reflect on feedback from peer review of course materials, for example, they probably derive the intended benefit of improving the quality of their teaching and student achievement (Berk, 2006). On the other hand, sometimes employing a measure leads to unintended consequences, such as when an instructor lowers standards and expectations for students based on the erroneous belief that doing so will lead to higher student ratings (Benton, Guo, Li, & Gross, 2013).

In considering the four aspects of trustworthiness, it quickly becomes apparent that not all sources of evidence "measure up." The reliability of well-designed SRI instruments exceeds that of other measures, because the resulting scores are based on the observations of multiple raters across multiple occasions. However, if an SRI is poorly designed, interpretations of its data may lack validity. Analyses from peer review of course materials, on the other hand, may be quite valid as long as colleagues take seriously their role of assisting in the improvement of their colleague's teaching performance (Braskamp & Ory, 1994; National Research Council, 2003). Nonetheless, they can be unreliable if peer raters differ markedly in their beliefs about what makes an effective teacher.

Distinguish Between Types of Evaluation

Once multiple measures have been collected, the evidence that is considered trustworthy should be reviewed to evaluate—i.e., render judgments about—the quality of teaching. Gathering such information typically has a threefold purpose: to guide teaching-improvement efforts, to determine how effectively the instructor taught, and to enhance student learning. The process of drawing conclusions about whether someone is teaching effectively or ineffectively is called *summative evaluation*. Because it is not a completely objective activity, decision makers should draw upon multiple sources of evidence, including students, the instructor, other faculty, and the administrator or employer (Berk, 2018). They should gather evidence for each course taught (although not necessarily for every term) and should examine trends in order to identify improvements or declines in teaching effectiveness (Hoyt & Pallett, 1999).

In contrast, when the accumulated evidence is examined for the purpose of making recommendations on how to improve teaching, educators are engaged in *formative evaluation*. Receiving feedback about "how things are going" in the classroom can change instructors' beliefs and attitudes about their students, strengthen confidence in teaching (Yi, 2012), and lead

to improvements in instructor performance (Andrade & Cizek, 2010), as well as increases in students' self-reported learning and satisfaction at the end of the term (Snooks, Neeley, & Revere, 2007). When responding to formative feedback, instructors should focus on one course at a time and on a limited number of teacher behaviors or course features (e.g., exams, active-learning strategies, course assignments; Buller, 2012; Hoyt & Pallett, 1999). Otherwise, with the abundance of courses taught, along with research activities and service responsibilities, college faculty can easily become overwhelmed and discouraged. Focusing on and seeing improvement in one area can, in contrast, strengthen motivation and teacher self-efficacy (Svinicki, 2017).

Although summative and formative evaluation have different purposes, it is difficult and perhaps unrealistic to keep them absolutely separate, because the information collected supports and informs both processes (Theall & Franklin, 2010). On IDEA's *Diagnostic Feedback* class report, for example, guidelines are provided to the instructor for conducting both types of evaluation. IDEA's global summary scores include average student self-ratings of progress on relevant objectives and ratings of the overall excellence of the teacher and the course, which serve the institution's needs for summative evaluation and accountability, whereas IDEA's diagnostic information about teaching methods can be used by faculty for formative purposes.

Consider the Benefits of Evaluation

The benefits of evaluating teaching are many. Summative evaluation aids in discerning which teaching approaches seem to be most and least effective, which courses a particular instructor is best prepared to teach, and which class sections students select. Administrators and faculty committees use evaluation evidence to decide whether to select certain teachers for an award, whether to retain or promote them, and how to assign merit-based pay increases. Formative evaluation can also be beneficial, especially when it reveals the exact nature of teaching difficulties, such as lack of clarity and conciseness. In order to learn and improve, instructors need specific feedback about where they have been successful and where they have fallen short (Andrade & Cizek, 2010; Hattie & Gan, 2011).

However, feedback alone may not necessarily reveal what specifically an instructor must do to modify teaching behaviors or remedy a problem. The faculty member—with the assistance of teaching and learning specialists, colleagues, and administrators—must consider many factors to devise a reasonable plan for addressing the problem. Evaluation can thus also advance the scholarship of teaching, because instructors respond to feedback by investigating new methods and strategies. Peers may even turn to one another for assistance and mentoring, thereby bolstering collegiality (Boyer, Moser, Ream, & Braxton, 2016; Cashin, 1996).

When consulting with a colleague, it is important to provide honest, realistic feedback about the instructor's capabilities. Although honest feedback can sometimes be hard to give and to receive, it can contribute to the development of a realistic view of oneself as a teacher (Hanna & Dettmer, 2003). Instructors need to know if they have weaknesses in certain areas; e.g., that they may be struggling with displaying personal interest in students or that their course organization is inadequate. This helps them monitor their progress as teachers.

The best kind of feedback leads to change. Formative feedback is most effective when it focuses on specific behaviors rather than on the person, when it is descriptive rather than evaluative, and when it occurs as soon as possible after performance (Brinko, 1993). Such information reveals the extent of teaching effectiveness, the required modifications in teaching methods, and the challenges that must be overcome in specific courses. Pellegrino, Chudowsky, and Glaser (2001, p. 87), as cited in Wilson and Scalise (2006, p. 636), explain:

> [A] major law of skill acquisition involves *knowledge of results.* Individuals acquire a skill much more rapidly if they receive feedback about the correctness of what they have done. If incorrect, they need to know the nature of their mistake. It was demonstrated long ago that practice without feedback produces little learning.

Feedback thus provides external guidance on how to regulate and monitor future teaching behavior. When

instructors observe a discrepancy between current and desired performance, it brings on cognitive dissonance, which can create a climate for change (Brinko, 1993). Teachers can then seek input from a trusted colleague or mentor to learn about specific actions that they can take. Without some type of consultation, feedback alone is unlikely to lead to improved instruction (Brinko, 1993; Cohen, 1980; Hampton & Reiser, 2004; Hativa, 2014; Knol, 2013; Marincovich, 1999; Marsh 2007; Marsh & Roche, 1993; Ory & Ryan, 2001; Penny and Coe, 2004).

Create a Growth Mind-Set

Unfortunately, however, what too often happens is that faculty are evaluated and ranked but not given sufficient feedback about how to improve. Most people probably respond negatively to feedback, especially when it comes only in the form of a numerical rating, with no recommendations on how to become a better teacher (Culbertson, Henning, & Payne, 2013). SRI, for example, are sometimes used almost like a Nielsen rating, merely judging teachers rather than providing input that would help them understand their strengths and weaknesses. Evaluation then becomes a demoralizing and dreaded experience, and student ratings, in particular, are despised. The result is that the ratings are tossed aside, and no systematic approach to professional development is applied. Discouragement follows when, year after year, the rankings among faculty within a unit remain largely unchanged. Such stability in rankings creates in faculty a fixed mind-set, which holds that skills, such as teaching ability, remain static throughout one's lifetime (Dweck, 2006).

An alternative is to adopt a growth mind-set, which holds that people can learn, grow, and better themselves (Dweck, 2006). By focusing on individual efforts and comparing an instructor's current progress to his or her past performance, rather than simply comparing one faculty member to another, administrators can help to foster a mind-set that focuses on developing better teaching methods and skills. Such an approach instills confidence, treats individuals fairly, and leads to growth and success (JCSEE, 2009).

Furthermore, a growth mind-set downplays performance rankings, a practice which usually makes everyone feel bad—except for the person doing the rankings (Rock, Davis, & Jones, 2014)! Some administrators may even derive satisfaction from reviewing and evaluating faculty materials and then placing instructors into categories. However, even the most highly ranked faculty members may feel deflated because of the limitations in funding available for merit-based salary increases. As for the rest of their departmental colleagues, most probably feel unappreciated and somewhat resentful. A collegial atmosphere is more likely when faculty help one another improve and accept constructive feedback about their work rather than compete for small stakes (Buller, 2012).

Adopting a growth mind-set, administrators can conduct in-depth conversations with individual faculty about personal goals, progress made toward those goals, and the contributions made to the department and institution (Rock et al., 2014). Then, "sitting beside" the instructor, the chair or head can discuss the results of the evaluation, and together they can chart a path to improved teaching. Personnel tend to respond well to such an approach because it is constructive, not demeaning or demoralizing. When done effectively, according to Braskamp and Ory (1994), it

- leads to discussion, planning, and action;
- addresses specific performance-enhancing behaviors;
- involves faculty in developing feedback strategies (ownership breeds acceptance);
- comes from multiple perspectives, which enhance credibility;
- finds patterns and commonalities among multiple perspectives;
- is trusting and supportive;
- rewards faculty who take the feedback seriously and respond with action.

Along these lines, Buller (2012) offers specific suggestions for how to give constructive feedback:

1. Begin by making it clear that you are trying to be helpful, not critical.
2. Cite specific, concrete examples of strengths and weaknesses in teaching behaviors. Be clear about which behaviors need to be

addressed and offer suggestions for improvements.
3. Allow the individual time to process the feedback, agreeing to meet another time in the near future for follow-up.
4. When offering praise and recognition, be as specific as possible. Cite examples of exceptional performance. Allow the faculty member time to enjoy the positive feedback.
5. Be sensitive to individual differences. Some faculty prefer private recognition, whereas others welcome public acclaim.

Best Practices in the Evaluation of Teaching

Of the three professional obligations placed upon college professors—commonly defined as teaching, research, and service—teaching is perhaps the most complex and difficult to evaluate. Care must therefore be taken to increase fairness and thoroughness in the evaluation process. The following elements, later described in detail, seem most essential for enhancing the effectiveness and usefulness of teaching evaluations.

- Recognize that the efficacy of an evaluation depends on the instructor.
- Create policies and procedures consistent with the institution.
- Uphold propriety standards.
- Clearly communicate expectations.
- Employ a balanced teaching-evaluation system.
- Include both formal and informal measures.
- Include authentic measures.
- Adopt a mastery orientation for formative evaluation.
- Adopt flexible evaluation schedules.
- Make the evaluation process useful.
- Ensure the accuracy of the system.
- Be sensitive to cultural and group differences.
- Use statistics appropriately.

Recognize that the Efficacy of an Evaluation Depends on the Instructor

The efficacy of an evaluation process rests with the instructor. Although the process of evaluation can itself foster self-reflection (Abbitt, 2011; Kim, 2011), the greatest gains come when the instructor actually does something with the feedback received. Marked improvements are consequently found among instructors who combine evaluation feedback and consultation with a colleague, mentor, or faculty-development specialist (Hampton & Reiser, 2004; Hativa, 2014; Knol, 2013; Marsh, 2007; Penny & Coe, 2004). What McKeachie (1976) said four decades ago is therefore true today—faculty are more likely to change if they gain insight from appropriate information sources (e.g., students, peers), are motivated to improve, and receive recommendations on how to improve.

Create Policies and Procedures Consistent with the Institution

Units that establish detailed procedures for preparing materials and conducting evaluations increase the likelihood that the process will be fair and meaningful. Written procedures must describe the information that should be gathered, what counts as acceptable evidence of teaching effectiveness, and all possible data sources. The American Association of University Professors (AAUP, 1975) specifically recommends that students be one source of such evidence.[1] Widespread acceptance and ownership of the written procedures is more likely when campus leaders and faculty are involved in the development of the evaluation system. Without the input of higher-level administrators, chairs run the risk of employing a system that contradicts institutional mission and goals (Cashin, 1996). The hours that faculty spend talking about the process, voicing their concerns, and holding meetings about evaluation are therefore not wasted; on the contrary, they increase the likelihood that the system will be fair and legal.

Uphold Propriety Standards

Of paramount importance when creating policies and procedures is to uphold propriety standards, conforming to ethical and legal principles, that protect those being evaluated (JCSEE, 2009). Leaders violate such standards when they fail to (a) align evaluation criteria with the mission and goals of the institution; (b) provide the resources necessary for faculty to improve;

[1] In 2005, the AAUP reaffirmed its commitment to the 1975 *Statement on Teaching Evaluation* in its *Observation on Association's 1975 Statement on Teaching Evaluation,* https://www.jstor.org/stable/40252838?seq=1#page_scan_tab_contents.

(c) recognize and reward excellent performance; and (d) address unsatisfactory performance in a timely manner. Academic units violate propriety standards when their policies are vague, unrealistic, or do not adhere to legal and ethical requirements or when they fail to communicate said policies to all employees. In upholding propriety standards, evaluators protect confidentiality, treat others with respect, and provide timely feedback that identifies strengths and maps out areas for growth.

Related to the propriety standard is the principle that individuals need to be educated in how to evaluate. Students, for example, can be instructed ahead of time on the value and purpose of the student-ratings system, the meaning of individual items, reasons that certain objectives were selected as relevant, and the importance of completing the ratings as requested (Ali, Al Ajmi, & Ali, 2013). Peers too can be educated in how to observe teaching and how to evaluate course materials. Administrators can also benefit by reading about how to evaluate teaching portfolios, discriminate among faculty on the basis of performance, and communicate feedback effectively.

Clearly Communicate Expectations

Because the relative importance of teaching, research, and service varies from one institution to the next, it is incumbent upon the administration to meaningfully convey its values and expectations regarding teaching performance (Paulsen, 2002). For example, if teaching is the highest priority in the institution, the procedures for evaluating and improving it ought to reflect its importance. Do unit standards communicate in advance what it means to be an excellent teacher? Do they state the expected teaching load? Do they specify which kind of data are acceptable as evidence of teaching effectiveness? Are there certain expectations for mean SRI scores? Much discussion among faculty and administrators is necessary (AAUP, 1975; Cashin, 1996; Seldin, 2006).

Such discussions may need to address whether expectations for faculty responsibilities and assignments can vary, depending on each individual's talents and interests (Paulsen, 2002). Some instructors excel, for example, with large sections of undergraduates, whereas others work well with upper-level students majoring in their subject area. Whereas some take a liking to designing and managing online courses, others prefer teaching face-to-face. One faculty member may have a talent for curriculum mapping, and another may connect well with students in a one-on-one advising role. Regardless of one's strengths, the department chair and each faculty member should meet at least annually to discuss and agree on expectations and the criteria that will be used to evaluate teaching performance.

A related question is, What constitutes meritorious teaching? One approach is to determine merit quantitatively, such as by the number of course sections taught, number of different courses developed, number of students advised, number of masters- and doctoral-level committees served, and size of course enrollments. Faculty members who agree to such a system communicate that they value all aspects of teaching, but that those who perform strongly across all areas of the teaching workload are deserving of the greatest merit.

In contrast, a qualitative approach holds that the quantity of work performed is not as important as its quality. Of what good is it to advise dozens of students poorly, teach multiple sections shoddily, and supervise several doctoral dissertations clumsily? Rather, performance should be recognized as meritorious when there is evidence of high student achievement and teaching excellence, such as awards and other types of recognition (Elmore, 2008).

Employ a Balanced Teaching-Evaluation System

Considerations of how to assign merit are complicated by the complexity of the teaching process itself. A balanced system of evaluation is needed, therefore, because of the limitations of any single measure (Berk, 2018; Lyde, Grieshaber, & Byrns, 2016; National Research Council, 2003). Student ratings should be just one leg of a three-legged stool that includes instructor self-assessments and review of course materials by relevant other parties (e.g., colleagues or peers, administrators, faculty-development specialists, external reviewers; Benton & Ryalls, 2016; Paulsen, 2002; Seldin, 2006).

Students, for example, are certainly qualified to offer a unique perspective on their own personal characteristics (e.g., background preparation, work

habits, motivation), student-teacher interactions, the instructor's classroom behavior, the perceived difficulty of the course, and their progress on learning objectives. However, others, such as peers, are more competent to assess an instructor's expertise and subject-matter knowledge; the quality of the course syllabus, readings, assignments, class materials, and exams; the appropriateness of course objectives; and the level of student achievement. Administrators, for their part, can appraise the instructor's scholarliness, contributions to the department, professionalism when interacting with students and colleagues, punctuality regarding grade submissions and other deadlines, and compliance with accreditation and curriculum alignment (Arreola, 2006; Braskamp & Ory, 1994).

Beyond those sources, instructors' self-assessments are also valuable, because they know best the planning that went into the course, the adjustments that they have made, and their personal growth as teachers. Although the validity of self-assessment can be compromised by instructor self-interest, self-evaluation nonetheless offers the benefit of fostering self-reflection and helping the instructor gain personal insight into how to improve (Braskamp & Ory, 1994; McGovern, 2006). It thus follows that instructors should present evidence of such reflection in the form of a teaching-philosophy statement, descriptions of course revisions and professional-development activities, class-time observations of student performance, and examples of questions that students have asked that indicate their quest for additional learning. Videotapes of actual instruction, samples of student work, and other evidence of student learning and development can also be added.

Units that take a balanced approach recognize the challenges in evaluating teaching effectiveness. The accumulated evidence must come from multiple sources and include materials such as descriptions of teaching activities, modifications made to courses, adoption of new teaching strategies, participation in professional-development activities, and contributions made to better the unit's overall instruction. Multiple measures increase the likelihood that the evaluation will encompass all dimensions of teaching, including course design, course delivery, course assessments, instructor availability, and course management (JCSEE, 2009).

Include Both Formal and Informal Measures

A balanced evaluation system also requires different types of measures. *Formal measures* evaluate teaching effectiveness relative to specific curriculum objectives, professional standards, or relevant benchmarks, such as comparisons with instructors in the same discipline or institution. One example is a standardized score from an SRI instrument, such as IDEA's comparative T-score for average student progress on relevant objectives. Another example would be a standard observational rubric for peer ratings, used during classroom visitation or for review of course materials.

Formal measures have several advantages. They are typically characterized by standardized administration procedures, uniform questions or rating scales, and consistent scoring systems, all of which enable comparisons among classes and instructors as well as longitudinal comparisons of the same instructor over time. The instruments are also typically backed by research and sometimes offer diagnostic feedback, which enables instructors to obtain specific information about their strengths and weaknesses.

However, formal measures have some disadvantages. For one, the questions are often worded too generally or are not suited to every type of course, especially practica, field experiences, labs, and courses in the performing arts. (Most systems, however, allow the instructor to add extra questions relevant to the specific course.) Another disadvantage is that although formal measures may identify areas of strength and weakness, they often do not provide guidelines for how to improve. However, the instructor can consult resources and colleagues to find the answers to that question.

Informal measures, which are recurring observations made by instructors during interactions with students, offset the disadvantages of formal measures. In practice, college faculty spend substantially more time interacting with students and observing them than they do administering formal assessments. Examples include observations of students as they work on small-group activities, complete problem sets, or hone skills; questions that students ask in class; and results of think-pair-share exercises. Such informal measures are nonstandardized, may or may not have scores, typically do not involve comparisons, are often characterized by

written comments based on observations or interviews, and are usually collected during class time, which makes them especially useful in formative evaluation.

Informal measures have several advantages. First, they are amenable to daily, continuous feedback about what is and is not working in the classroom. Second, instructors can make corrections "on the fly" instead of waiting until a fixed point in time, when a formal assessment is administered. Third, informal assessment can either confirm or contradict what is learned from the information collected using a formal instrument (Ormrod, 2014). Whereas the formal measure may identify the problem, informal assessment may better help determine the solution.

Although the idea of informal measures is not new, the case for embedded measures has been made more recently, based on their increasing relevance to the evaluation of teaching (Benton & Li, 2015; Wieman, 2015). Embedded measures, collected during class, provide evidence of students' progress toward learning outcomes, such as their performance on activities, assignments, projects, and papers. By connecting accomplishments with specific learning outcomes, the instructor can assess whether students are grasping the important concepts and whether teaching modifications should be made.

Even so, informal measures lack standardization, which prevents comparisons among instructors or across courses. In addition, evidence for validity is often lacking due to discrepancies between what instructors observe and what students actually experience, and reliability can be low due to inconsistency in the accuracy of what is recalled or recorded (Robinson & Lubienski, 2011). For that reason, it is wise to keep written records (Miller, Linn, & Gronlund, 2009; Stiggins & Chappuis, 2012).

Include Authentic Measures

When assembling evidence of teaching effectiveness, instructors should consider including authentic measures, which involve observing, interpreting, and synthesizing information derived from students' performance in realistic situations. Instructors can collect evidence of both the outcomes of authentic activities and their processes. For example, evidence might include the products of student work (e.g., written papers, studio designs), as well as recordings of student presentations.

Authentic measures are necessary for several reasons. First, they can inform instructors about students' needs for additional instruction, because when students perform an authentic task, their need for additional knowledge and skills is revealed. The opportunity for improved teaching thus presents itself. Second, authentic measures can demonstrate to colleagues the level of student achievement of learning outcomes. Rather than relying only upon student test scores or self-reported progress scores from a single SRI instrument, the instructor can encourage students to track their progress across time using a number of authentic measures. These measures can be compiled in student portfolios, which contain (a) selected samples of students' work and (b) records that document growth and status in a content domain. Such portfolio assessment better represents students' achievements than does a single test score. For more information on student learning portfolios and how they can be used in formative evaluation of instruction, see Zubizarreta (2008).

A third reason to collect authentic measures is that they reveal what students can actually *do*. In contrast, more traditional measures, such as tests, tend to emphasize only students' recall of what they learned in the class or, as in the case of student ratings, what students perceive they achieved and the behaviors that they perceive their teacher performed. Such measures, however, offer little insight into how students handle real-life problems, and actions that instructors might take to improve instruction toward that end.

Adopt a Mastery Orientation for Formative Evaluation

Trying out new teaching methods, such as integrating student active-learning strategies into a course, is risky. What if students react negatively or uncooperatively? How will that affect the instructor's end-of-course student ratings? When instructors implement such changes in response to formative evaluation, it is best to adopt a mastery-goal orientation, which focuses on learning, risk taking, openness to feedback for improvement, and persistence (Elliot & Harackiewicz, 1996; Svinicki, 2016.)

For example, a mastery approach would encourage the instructor not to worry about making mistakes (i.e., see "Fear of Looking Stupid"). She could then be encouraged to seek help from colleagues who have more experience, to put in the necessary time and effort, and to accept responsibility for the results. Regardless of whether student outcomes increase in the first trial semester, the instructor can be pleased with her effort rather than fear the possibility of low student ratings. A mastery orientation is especially beneficial for new faculty who are just learning how to teach effectively. They can test new approaches, get feedback from students and peers, and work to improve their performance.

In contrast, a performance-goal orientation evaluates teaching by comparing one's performance to others' (Elliot & Harackiewicz, 1996; Svinicki, 2016). Judgments about how well "I" do then depend upon how well everyone else does. Such an approach usually creates an atmosphere of competition among colleagues because, rather than cooperating with one another, faculty compete for a position in the performance rankings. As a consequence, highly rated teachers sometimes develop false perceptions of their abilities, whereas those consistently ranked at the low end may become discouraged and relatively less motivated to make changes.

The distinction between mastery-goal and performance-goal orientations has parallels to the fixed versus growth mind-sets mentioned previously. When the focus is on comparisons with others (performance-goal orientation) rather than on individual progress (mastery-goal orientation), instructors may develop a fixed mind-set, which can lead to the adoption of a performance-avoidance orientation (Svinicki, 2016). This behavior occurs when, to avoid appearing incompetent, instructors choose not to try out new teaching strategies, for fear of failure, or decide not to seek help, for fear of admitting to shortcomings in their teaching. This orientation leads to anxiety, a reluctance to take risks, and limits on how much instructors learn.

But effective evaluation should lead instructors to take risks! It should encourage persistence in the aftermath of mistakes and experimentation. Enhanced student learning is therefore best achieved by taking a mastery approach to formative evaluation, which should lead to increased faculty vitality. With mastery goals, "not everyone need be judged by the same yardstick," and consideration is given to diverse career paths and stages of development (Buller, 2012, p. 68).

Adopt Flexible Evaluation Schedules

Career paths and development also diverge in the teaching assignments and loads that faculty carry. Accordingly, evaluation schedules must be flexible enough to meet the diverse needs and requirements for first-year (nontenured), adjunct, other nontenured, and tenured faculty (Hoyt & Pallett, 1999). For first-year faculty, it is in the institution's best interest to give attention to each course taught, as there may be need for improvements and decisions about course assignments and merit-pay recommendations. It may also be advantageous to evaluate experienced adjunct faculty and other nontenured faculty annually for salary recommendations and, in the case of adjuncts, decisions about retention. When the time comes for tenure or promotion decisions, at least two sets of each data source (e.g., SRI, peer reviews of course materials, instructor self-assessments), one from early in tenure and the other more recent, are recommended for each course taught. Aside from the need for summative evaluations for annual salary recommendations, formative evaluation of tenured faculty might be conducted at least once every few years for every course taught.

Make the Evaluation Process Useful

From a practical standpoint, evaluation works best when it helps instructors improve their teaching performance. Useful evaluation leads to meaningful professional development—the utility standard of evaluation (JCSEE, 2009). Administrators fall short of this standard when they fail to provide written reports of an evaluation in a timely manner, naively assume that all raters and instructors will use the information contained in the evaluation appropriately, fail to respond to the instructor's strengths and weaknesses, set unreasonable expectations or expectations inconsistent with the institution's goals and mission, fail to nurture a climate of support and growth, and fail to follow up to see what improvements have been made. Useful evaluation is also connected to the reward structure, such that highly effective teachers are rewarded differently than less effective ones.

Ensure the Accuracy of the System

To be useful, the system must produce sound information; otherwise what is the value of participating in the evaluation in the first place? When evaluators use instruments with faulty reliability or validity or use them for other than their intended purpose, when they focus only on what is easiest to measure or count, and when they fail to take into consideration factors beyond the instructor's control (e.g., student motivation, class size), precision is reduced (Benton & Ryalls, 2016; Narayanan, Sawaya, & Johnson, 2014). In addition, an accurate evaluation system requires neither underestimating nor overestimating the influence of contextual factors on interpretations of performance, such as culture, gender, course delivery, and academic discipline (Young & Duncan, 2014; Young, Rush, & Shaw, 2009). Whereas an overly complex system can render evaluation impractical—as in trying to cut a log with a razor (Pallett, 2006)—accuracy suffers when the process is too informal. Evidence for the reliability and validity of all measures therefore reinforces confidence in the system. Moreover, occasional reviews of the evaluation process help spot necessary revisions in light of recognized shortcomings.

Be Sensative to Cultural and Group Differences

Effective evaluation is sensitive to cultural and group differences, such as gender, age, ethnicity, disability, religion, nation of origin, socioeconomic status, and sexual orientation (JCSEE, 2009). Evaluators must always be aware of their own biases and assumptions and sensitive to the worldviews of others. Racial and gender biases cut across all facets of society, and a fair evaluation system requires an examination of the extent to which they exist at an institution or within an academic unit (Baldwin & Blattner, 2003). Bias also likely exists to some degree in any evaluation system, because it is designed and carried out by humans. All sources of evidence—student ratings, peer observations, instructor self-assessments, and peer/supervisor reviews—are therefore subject to bias. But evaluation can still be useful, provided the system is well designed, multiple sources of evidence are submitted, and possible sources of bias are recognized.

Use Statistics Appropriately

Most formal assessments produce reports that contain scores based on some statistical computations. Sometimes statistical analyses can be usefully applied to control for extraneous influences due to group and individual differences (e.g., class size, student preparedness, student level). In the IDEA SRI system, for instance, separate "norms" are computed for different academic disciplines, and adjusted scores control for differences in average student motivation to take the course, background preparation, and work habits (Li, Benton, Brown, Sullivan, & Ryalls, 2016). But at other times, statistics are overemphasized, especially when evaluators make too much of too little (Pallett, 2006). Administrators and faculty, for example, may sometimes make judgments about teaching effectiveness based on small differences in mean student-ratings scores (Boysen, Kelly, Raesly, & Casner, 2014). But, as McKeachie points out (2007, p. 465), decision makers draw an erroneous conclusion when they compare scores between instructors and assume, just because "one number is larger than another, there is a real difference between the teachers to whom the numbers have been assigned." When this happens, "faculty members whose students 'agree' that they are excellent teachers may find that they fall 'below average' in comparisons with those whose students 'strongly agree' on that same item."

By taking a balanced approach to evaluation, which incorporates multiple sources of evidence, such trivial differences in scores are less likely to guide decision making. But even on campuses where student ratings are used as the primary or only evidence of teaching effectiveness, the following are good practices (Buller, 2012).

1. Look beyond mean raw scores. In the IDEA SRI system, for example, raw scores are converted to T-scores that express an individual's rating in standard-deviation units above or below the mean of 50. The system then places each faculty member's score into one of five categories, ranging from *Much Higher* to *Much Lower.* Examining standard deviations and frequencies for each item can also reveal the spread in scores and possible challenges that the instructor faced if the students were particularly heterogeneous in their views and experiences.

2. Consider contextual factors and student and course characteristics beyond the instructor's control. For example, student ratings vary by discipline. Courses in science, technology, engineering, and mathematics (STEM) tend to be rated less highly than those in non-STEM fields (Benton, 2015; Li & Benton, 2017). In addition, ratings are somewhat lower in large classes than in small ones and in courses with students who report insufficient background preparation, less motivation to take the course, and poorer work habits (Li et al., 2016). Institutions might also assess whether ratings differ between online and face-to-face courses, as has sometimes been the case (Young & Duncan, 2014), although others have found more similarities than differences (Benton, Webster, Gross, & Pallett, 2010; McGhee & Lowell, 2003; Wang & Newlin, 2000).
3. Look for patterns for the same instructor across courses. However, simply calculating the mean of the means across courses is misleading, because each course rating is based on a different number of students, and extreme scores can either inflate or depress the average. Instead, try to determine if students consistently communicate the same perceptions of their experiences with an instructor.

Conclusions

When done effectively, evaluation of teaching can create a growth mind-set that instills confidence in instructors. Rather than being demoralized by performance rankings that create a fixed mind-set, faculty can concentrate on their individual efforts and compare their current progress to past performance. They can then focus on developing better teaching methods and skills, rather than fearing or resenting comparisons to others. Effective evaluation is complex and requires the use of multiple measures—formal and informal, traditional and authentic—as part of a balanced evaluation system. The student voice, a critical element of that balanced system, is appropriately complemented by instructor self-assessment and the reasoned judgments of relevant other parties, such as peers and supervisors. Integrating all three elements allows instructors to take a mastery approach to formative evaluation, trying out new teaching strategies and remaining open to feedback that focuses on how they might improve. Such feedback is most useful when it occurs within an environment that fosters challenge, support, and growth. By taking these steps, evaluation of teaching becomes a rewarding process, not a dreaded event.

Steve Benton is senior research officer at the IDEA Center where, since 2008, he has led a research team that designs and conducts reliability and validity studies for IDEA products. He received his PhD in psychological and cultural studies at the University of Nebraska–Lincoln in 1983. He is a Fellow in the American Psychological Association and American Educational Research Association, as well as an emeritus professor and former chair of Special Education, Counseling, and Student Affairs at Kansas State University, where he served for 25 years. His current research focuses on best practices in faculty development and evaluation.

Suzanne Young is a professor of educational research and associate dean for graduate programs in the College of Education at the University of Wyoming. She received her PhD from the University of Northern Colorado in 1995, majoring in educational psychology with an emphasis on research, statistics, and assessment. Her research interests include work in the areas of effective teaching in higher education, online teaching and learning, and student engagement and classroom community in distance environments.

IDEA Papers authored or co-authored by IDEA staff are edited by non-staff editorial board members

References

Abbitt, J. T. (2011). An investigation of the relationship between self-efficacy beliefs about technology integration and technological pedagogical content knowledge (TPACK) among preservice teachers. Journal of Digital Learning in Teacher Education, 27(4), 134–143. doi:10.1080/21532974.2011.10784670

Ali, H. I. H., & Al Ajmi, A., Ali Saleh. (2013). Exploring non-instructional factors in student evaluations. Higher Education Studies, 3(5), 81–93. doi:10.5539/hes.v3n5p81
American Association of University Professors. (1975). Statement on teaching evaluation. https://www.aaup.org/report/statement-teaching-evaluation

American Educational Research Association, American Psychological Association, National Council on Assessment in Education (2014). Standards for educational and psychological testing. American Educational Research Assn.

Andrade, H. L., & Cizek, G. J. (Eds.). (2010). Handbook of formative assessment. New York: Routledge.

Arreola, R. A. (2006). Developing a comprehensive faculty evaluation system (2nd edition). Bolton, MA: Anker Publishing.

Baldwin, T., & Blattner, N. (2003). Guarding against potential bias in student evaluations. College Teaching, 51, 27–32.

Bana e Costa, C. A., & Oliveira, M. D. (2012). A multicriteria decision analysis model for faculty evaluation. Omega, 40, 424–426.

Benton, S. (2015). Student ratings of instruction in lower-level postsecondary STEM classes (pp. 59–72). In Searching for better approaches: Effective evaluation of teaching and learning in STEM. Tucson, AZ: Research Corporation for Science Advancement.

Benton, S. L., & Cashin, W. E. (2011). IDEA Paper No. 50: Student ratings of teaching: A summary of research and literature. Manhattan, KS: The IDEA Center. http://theideacenter.org/research-and-papers/idea-papers/50-student-ratings-teaching-summary-research-and-literature

Benton, S. L., Guo, M., Li, D., & Gross, A. (2013, April). *Student ratings, teacher standards, and critical thinking skills*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Benton, S. L., & Li, D. (2015). Response to "A better way to evaluate undergraduate teaching. *IDEA Editorial Note #1*. Manhattan, KS: The IDEA Center. http://ideaedu.org/research-and-papers/editorial-notes/response-to-wieman/

Benton, S. L., & Ryalls, K. R. (2016). *IDEA Paper #58: Challenging misconceptions about student ratings of instruction*. Manhattan, KS: The IDEA Center.

Benton, S. L., Webster, R., Gross, A. B., Pallett, W. (2010). *IDEA Technical Report No. 15: An analysis of IDEA Student Ratings of Instruction in traditional versus online courses.* Manhattan, KS: The IDEA Center.

Berk, R. A. (2006). *Thirteen strategies to measure college teaching*. Sterling, VA: Stylus.

Berk, R. A. (2018). Start spreading the news: Use multiple sources of evidence to evaluate teaching. *Berk's Law, 32*, 73–81.

Boyer, E. L., Moser, D., Ream, T. C., & Braxton, J. M. (2016). *Scholarship reconsidered: Priorities of the professoriate*. San Francisco, CA: Jossey-Bass.

Boysen, G. A., Kelly, T. J., Raesly, H. N., & Casner, R. W. (2014). The (mis)interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education, 39*(6), 641–656. doi:10.1080/02602938.2013.860950

Bradley, K. D., & Bradley, J. W. (2010). Exploring the reliability, validity, and utility of a higher education faculty review process. *Contemporary Issues in Education Research, 3*, 21–26.

Braskamp, L. A, & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performance.* San Francisco: Jossey-Bass.

Brinko, K. T. (1993). The practice of giving feedback to improve teaching: What is effective?" *Journal of Higher Education, 64*, 54–68.

Buller, J. L. (2012). *Best practices in faculty evaluation*. San Francisco, CA: Jossey-Bass.

Cashin, W. E. (1996). *IDEA Paper #33: Developing an effective faculty evaluation system*. Manhattan, KS: The IDEA Center.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education, 13*, 321–341.

Culbertson, S., Henning, J. B., & Payne, S. C. (2013). Performance appraisal satisfaction: The role of feedback and goal orientation. Journal of Personnel Psychology, 12, 189–95.

Dweck, D. S. (2006). *Mindset*: *The New Psychology of Success: How We Can Learn to Fulfill Our Potential.* New York, NY: Ballantine Books.

Elliot, A. J., & Harackiewicz J. (1996). Approach and avoidance: Achievement goals and intrinsic motivation. Journal of Personality and Social Psychology, 70(3), 451–475.

Elmore, H. W. (2008). Toward objectivity in faculty evaluation. *Academe, 94*, 38–40.

Hampton, S. E., & Reiser, R. A. (2004). Effects of a theory-based feedback and consultation process on instruction and learning in college classrooms. *Research in Higher Education, 45*, 497–527.

Hanna, G. S., & Dettmer, P. S. (2004). *Assessment for teaching. Using context-adaptive planning*. Upper Saddle River, NJ: Pearson/Allyn and Bacon.

Hativa, N. (2014). Student ratings of instruction: Recognizing effective teaching. Oron Publications.

Hattie, J., & Gan, M. (2011). Instruction based on feedback. In R. E. Mayer & P. A. Alexander (Eds.), Handbook of research on learning and instruction (pp. 249–271). New York: Routledge.

Hoyt, D. P., & Pallett, W. H. (1999). IDEA Paper No. 36, Appraising teaching effectiveness: Beyond student ratings. Manhattan, KS: The IDEA Center. Retrieved from: http://ideaedu.org/wp-content/uploads/2014/11/idea_paper_36.pdf

Joint Committee on Standards for Educational Evaluation (2009). The personnel evaluation standards (A. R. Gullickson, Chair). Thousand Oaks, CA: Corwin Press.

Kim, Y. H. (2011). Prospective early childhood educators' meta-cognitive knowledge and teacher self-efficacy: Exploring domain-specific associations. Educational Psychology, 31(6), 707–721. doi:10.1080/01443410.2011.599924

Knol, M. (2013). Improving university lectures with feedback and consultation. Academisch Proefschrift. Ipskamp Drukkers, B.V.

Li, D., Benton, S. L., Brown, R., Sullivan, P., & Ryalls, K. R. (2016). IDEA Technical Report No. 19: Analysis of student ratings of instruction system 2015 pilot data. Manhattan, KS: The IDEA Center.

Li, D., Benton, S. L. (2017). IDEA Research Report #10: Examining instructor-gender-by-academic-discipline interactions in student ratings of instruction. Manuscript in preparation.

Lyde, A. R., Grieshaber, D. C., & Byrns, G. (2016). Faculty teaching performance: Perceptions of a multi-source method for evaluation (MME). Journal of the Scholarship of Teaching and Learning, 16(3), 82–94. doi:10.14434/josotl.v16i3.18145

Marincovich, M. (1999). Using student feedback to improve teaching. In P. Seldin, & Associates, Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions (pp. 45–69). Bolton, MA: Anker.

Marsh, H. W. (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.

Marsh, H. W., & Roche, L. A. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. American Educational Research Journal, 30, 217–251.

McGhee, D. E., & Lowell, N. (2003). Psychometric properties of student ratings of instruction in online and on-campus courses. In T. D. Johnson & D. L. Sorenson (Eds.), Online student ratings of instruction: New Directions for Teaching and Learning, No. 96 (pp. 39–48). San Francisco: Jossey-Bass.

McGovern, T. (2006). Self-evaluation: Composing an academic life narrative. In P. Seldin (Ed.), *Evaluating faculty performance: A practical guide to assessing, teaching, research, and service* (pp. 96–110). Bolton, MA: Anker Publishing Co., Inc.

McKeachie, W. J. (1976). Psychology in America's bicentennial year. *American Psychologist, 31*, 819–833.

McKeachie, W. J. (2007). Good teaching makes a difference and we know what it is. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based approach* (pp. 457–474). Dordrecht, The Netherlands: Springer.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Assessment and assessment in teaching* (10th ed.). Upper Saddle River, NJ: Merrill/Pearson.

Narayanan, A., Sawaya, W. J., & Johnson, M. D. (2014). Analysis of differences in nonteaching factors influencing student evaluation of teaching between engineering and business classrooms. *Decision Sciences Journal of Innovative Education, 12*(3), 233–265. doi:10.1111/dsji.12035

National Research Council (2003). *Evaluating and improving undergraduate teaching in science, technology, engineering, and mathematics.* Washington, D. C.: The National Academies Press.

Ormrod, J. E. (2014). Educational psychology: Developing learners. Upper Saddle River, NJ: Pearson.

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? In T. D. Johnson & D. L. Sorenson (Eds.), *Online student ratings of instruction: New Directions for Teaching and Learning*, No. (pp. 27–44). San Francisco: Jossey-Bass.

Pallett, W. (2006). Uses and abuses of student ratings. In P. Seldin, *Evaluating faculty performance* (pp. 50-65). Bolton, MA: Anker Publishing Company.

Paulsen, M. B. (2002). Evaluating teaching performance. *New Directions for Institutional Research, 114*, 5–18.

Penny, A. R., & Coe, R. (2004). Effectiveness of consultation on student ratings feedback: Meta-analysis. *Review of Educational Research,* 74, 215–253.

Robinson, J. P. & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal, 48*, 268–301.

Rock, D., Davis, J., & Jones, B. (2014). Kill your performance rankings. *Strategy+Business Magazine, 76*, 1–10.

Seldin, P. (2006). Building a successful evaluation program. In P. Seldin (Ed.), *Evaluating faculty performance: A practical guide to assessing teaching, research, and service* (pp. 1–19). Bolton, MA: Anker Publishing Co., Inc.

Snooks, M. K., Neeley, S. E., & Revere, L. (2007). Midterm student feedback: Results of a pilot study. *Journal on Excellence in College Teaching, 18*(3), 55–73. http://celt.miamioh.edu/ject/

Stiggins, R. J., & Chappuis, J. (2012). *An introduction to student-involved assessment FOR learning* (6th ed.). Boston: Pearson Assessment Training Institute.

Svinicki, M. D. (2016). *IDEA Paper #59: Motivation: An updated analysis*. Manhattan, KS: The IDEA Center.

Svinicki, M. D. (2017). From Keller's MVP model to faculty development practice. *New Directions for Teaching and Learning, 152,* 79–89.

Svinicki, M., & McKeachie, W. J. (2014). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers* (14th ed.). Belmont, CA: Wadsworth.

Theall, M. (2017). MVP and Faculty Evaluation. *New Directions for Teaching and Learning, 2017*: 91–98. doi: 10.1002/tl.20271

Theall, M. & Franklin, J. L. (2010). Assessing teaching practices and effectiveness or formative purposes. In K. J. Gillespie, D. L. Robertson, and Associates (Eds.), *A guide to faculty development* (pp. 151–168). San Francisco, CA: Jossey-Bass.

Wang, A. Y., & Newlin, M. H. (2000). Characteristics of students who enroll and succeed in psychology web-based classes. *Journal of Educational Psychology, 92,* 137–143.

Wieman, C. (2015). A better way to evaluate undergraduate teaching. Change, (Jan/Feb), 7–15.

Wilson, M., & Scalise, K. (2006). Assessment to improve learning in higher education: The BEAR assessment system. *Higher Education, 52*, 635–663.

Yi, Q. (2012). Empirical study of formative evaluation in adult ESL teaching. *English Language Teaching, 5*(2), 27–38. Retrieved from http://search.proquest.com.er.lib.k-state.edu/docview/1773221272?accountid=11789

Young, S., Rush, L. S., & Shaw, D. G. (2009). Evaluating gender bias in ratings of university instructors' teaching effectiveness. *International Journal for the Scholarship of Teaching and Learning. 3*(2), doi: 10.20429/ijsotl.2009.030219

Young, S. & Duncan, H. (2014). Online and on-campus teaching: How do student ratings differ? *Journal of Online Learning and Teaching, 10*(1), 70–79.

Zubizarreta, J. (2008). *IDEA Paper #44: The learning portfolio: A powerful idea for significant learning.* Manhattan, KS: The IDEA Center.

T: 800.255.2757
T: 785.320.2400

301 South Fourth St., Suite 200
Manhattan, KS 66502-6209
E: info@IDEAedu.org
IDEAedu.org