IDEA Editorial Note #1 • Response to "A Better Way to Evaluate Undergraduate Teaching"

Stephen L. Benton and Dan Li • The IDEA Center

In the January/February issue of *Change*, Carl Wieman (2015) argues, in "A Better Way to Evaluate Undergraduate Teaching," the current methods of evaluating teaching effectiveness in higher education lack validity and do little to offer the means for improvement. He proposes an alternative based on an inventory of instructor self-reported teaching practices, which are suggested by education research to be associated with student learning. As researchers for a non-profit organization that seeks to improve learning in higher education through research, assessment and professional development, we are delighted to see more research-driven efforts on developing instruments for teaching evaluation. Nevertheless, we regretfully find Wieman's critique of student ratings of instruction (SRIs), which he refers to as "student course evaluations," to be a limited perspective. In this article we review and respond to Wieman's critique, based on extensive research on validity and reliability of student ratings. To contribute to the mission of designing better ways to evaluate teaching, we also introduce a comprehensive system of faculty evaluation—widely supported by research and users—that includes SRIs as one of multiple measures of teaching effectiveness. In this system, evaluation is no longer considered just an end-of-a-course routine that instructors have to perform, but instead seen as an ongoing developmental process that requires triangulation of multiple sources of evidence.

RESPONSE TO WIEMAN'S VIEW ON TEACHING EFFECTIVENESS

Wieman's definition of teaching effectiveness, "producing the desired learning outcomes for the given student population" (p. 8), conforms to the underlying model of the IDEA Student Ratings of Instruction system, which was created by Donald P. Hoyt four decades ago (1973). More specifically, IDEA has always considered student progress on instructor-identified learning objectives the best measure of teaching effectiveness. However, we are reluctant to conclude the instructor "produces" the learning outcome, which is contrary to decades of research on the role of the student in learning. Students construct knowledge and are, therefore, active and responsible for their own learning. Moreover, students tend to give instructors who expect students to take the share of responsibility for learning higher ratings (Benton, Guo, Li, & Gross, 2013).

According to Wieman, student attainment of learning outcomes "is highly dependent on the backgrounds of the students" (p. 8). We agree somewhat. Yes, it is true that student backgrounds influence learning, as supported by decades of research in educational psychology. However, student ratings of their background preparation account for only a small fraction of their reported progress on course objectives emphasized by the instructor (Benton & Li, 2015). Other student characteristics are just as important—desire to take the course regardless of who taught it (motivation) and the typical effort they put forth on academic work (work habits).

We also agree "meaningful measures of teaching quality must separate out the impact of the teacher" (p. 8). This is why IDEA has always used adjusted scores, which control for factors that are beyond instructors' control but can influence student ratings. Based on research, IDEA employs a sophisticated scheme to single out instructors' impact on learning: It first removes the effects of instructors with respect to how much they stimulate students' intellectual effort and the amount of coursework, then adjusts for student background, effort, and difficulty of subject matter (Hoyt & Lee, 2002).

Wieman suggests the undervaluing of teaching at research universities may, in part, be attributed to the relative inferiority of existing measures of teaching effectiveness. We argue the current *practice* of evaluating teaching may be what leaves a lot to be desired. In contrast to Wieman's statement, quality measures of teaching effectiveness exist. Unfortunately, decisions are too often made on the basis of a single criterion. It would be as if research were judged by only one criterion—number of publications or amount of funding acquired. Neither by itself would signal quality research any more than an average student ratings score should be used as the only measure of teaching effectiveness.

REVIEW OF WIEMAN'S CRITERIA FOR JUDGING TEACHING EFFECTIVENESS MEASURES

Wieman describes five criteria in judging the quality of teaching-effectiveness measures. In the following we describe how IDEA meets each criterion, and we propose multiple criteria in addition to Wieman's.

Validity

Validity concerns whether evidence supports the interpretation of a test or score for its intended purpose. Any criterion of teaching effectiveness, Wieman argues, must be correlated with the achievement of the desired student outcomes. Across multiple sections of the same course taught by the same instructor, IDEA student ratings of progress on relevant course objectives are positively correlated with exam scores, whereas ratings on irrelevant objectives are not. Students who rate their progress as either exceptional or substantial generally outperform those reporting moderate or less progress on course examinations (Benton, Duchon, & Pallett, 2013). Multiple studies, using different instruments and measures of learning, have also found that student ratings of instruction do correlate positively with student achievement (see Benton & Cashin, 2014, for a review). The correlations typically range from .30 to .50, which is impressive given the restricted range of most student ratings scales (e.g., 5-point scale) and the uncertain validity and reliability of most classroom assessment. Moreover, instructors are neither the only cause of student learning nor the most important one (Hativa, 2013b). One would therefore not expect student ratings of the instructor to correlate highly with how much they learn in a given course.

Meaningful comparisons

To judge teaching effectiveness requires comparing individual performance to a standard. Such comparison provides feedback about how an instructor performs relative to others and about areas that need improvement. We couldn't agree more. For years IDEA has provided standard T-scores, which allow direct comparisons between items with different means and standard deviations, on the instructor's average score for student ratings of progress on relevant objectives (PRO) and global ratings of the teacher and the course. IDEA also provides norms in the discipline, the institution, and overall IDEA population of classes, which cuts across all Carnegie classifications, regions of the U.S., and public and private institutions.

Fairness

Wieman's notion of fairness is that a measure of teaching quality must be valid across all types of courses and instructors. He mentions several extraneous factors that can threaten such validity—class size, student level, subject matter, student preparedness, and institution type. At IDEA we recognize many factors beyond the instructor's control that can affect learning and ratings. Accordingly, we have adjusted scores for class size, student motivation and work habits (which are related to student level), and we will soon adjust for background preparation. As already mentioned, we provide separate comparison scores for the discipline and the institution. By controlling for these extraneous variables, we try to "level the playing field" among courses that vary along multiple dimensions.

Practicality

Practicality concerns matters of time and money. The fact that student ratings continue to be used on most campuses indicates their practicality. The typically high response rates of the students surveyed, and the relatively low cost per class for conducting ratings, suggest that student ratings are a practical, if not the most practical, approach to obtain feedback about instruction. Powered by mobile devices and automatic response systems, faculty can obtain instant feedback multiple times while the course is being conducted, so that they can measure the effects of their adjustments in response to previous feedback. What other system provides such immediate feedback from multiple observers at a relatively low cost?

Improvement

A quality measure of teaching effectiveness must provide clear guidance to instructors "not only on how well they are doing but how they can improve" (p. 9). We absolutely agree. When Donald P. Hoyt began creating a student ratings of instruction instrument in 1968, the first problem he and a representative group of faculty encountered was determining the purpose of the instrument. After much discussion, there was a consensus that "improving teaching effectiveness" should be the major focus (Hoyt, 1973), and it remains so today at IDEA.

WHAT'S MISSING IN WIEMAN'S CRITERIA?

Noticeably missing in Wieman's list is *reliability*. Reliability is important for determining whether a measure is consistent enough to be used as a source of evidence for making decisions about teaching effectiveness. If the measure varied substantially for the same instructor across the same course, judgments about teacher quality would be difficult. As it turns out, student ratings are actually the most reliable single measure of teaching effectiveness because they represent the observations of multiple raters across multiple class sessions (Marsh, 2007). No other measure can make that claim. IDEA has high reliability for the same instructor across multiple courses, especially when at least five classes have been rated (Benton & Li, 2015).

Another missing element of Wieman's criteria is that validity should always be tied to proper use of the measure. Unfortunately, SRIs are often overemphasized in summative evaluation and underutilized in formative evaluation. They are overemphasized when faculty and administrators rely

on them exclusively for evidence of teaching effectiveness in decisions about tenure, promotion, and merit salary adjustments. As pointed out in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014),

> Validation is the joint responsibility of the test developer and the test user. The test developer is responsible for furnishing relevant evidence and a rationale in support of any test score interpretations for specified uses intended by the developer. The test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used. When a test user proposes an interpretation or use of test scores that differs from those supported by the test developer, the responsibility for providing validity evidence in support of that interpretation for the specified use is the responsibility of the user (p. 13).

Wieman chose to rely upon only one aspect of validity evidence—correlation with the desired student-learning outcome—but multiple forms of evidence can demonstrate validity. Strong evidence in support of one interpretation of a measure (i.e., its relation to desired student-learning outcomes) "in no way diminishes the need for evidence to support other parts of the interpretation (AERA et al., 2014, p. 13). There is, for example, the need to determine whether SRIs represent a single dimension or multiple dimensions of teaching effectiveness. Substantial evidence has been found to support the multidimensionality of SRIs. Across several factor-analytic studies, SRIs have consistently shown an internal structure comprised of multiple underlying constructs (see Hativa, 2013a for a review). That is, SRIs are not just a unitary measure of how popular or attractive students find the instructor to be. On the contrary, SRIs tend to be comprised of four major areas of teaching effectiveness: organization, clarity, enthusiasm/expression, and rapport/ interactions (Hativa, 2013a).

Another important validity question concerns whether SRIs are related to other measures of teaching effectiveness. If not, then one would question their validity as a criterion measure. Substantial evidence has shown SRIs do correlate with other measures. SRIs are positively correlated with ratings of an instructor by the instructor, by colleagues and administrators, by alumni, and by trained observers (see Benton & Cashin, 2014, for a review). Relationships between SRIs and such relevant measures provide evidence for convergent validity. Divergent validity is demonstrated by low correlations between SRIs and variables purportedly unrelated to teaching effectiveness. For example, SRIs are weakly correlated with instructor age, gender, race, personal characteristics, and teaching experience. They are also unrelated to student age, gender, level, grade point average, and personality. Administration of SRIs, including the time of day and the time during the term, is irrelevant with the ratings (see Benton & Cashin, 2014, for a review).

Of further practical value is validity evidence demonstrating SRIs can actually lead to improvements in teaching. Supporting evidence is found in multiple studies showing positive effects of SRIs on teaching improvement, especially when they are combined with consultation (Brinko, 1990; Hampton & Reiser, 2004; Hativa, 2013a; Knol, 2013; Marincovich, 1999; Marsh 2007; Marsh & Roche, 1993; Ory & Ryan, 2001; Penny and Coe, 2004).

RESPONSE TO CRITIQUE OF STUDENT RATINGS (I.E., STUDENT COURSE EVALUATIONS)

We first take issue with Wieman's use of the term "student course evaluations." While such a usage is not uncommon, "evaluation" implies judgment of worth and requires relevant expertise or credentials. Faculty and administrators, rather than students, are the ones qualified to interpret SRIs (i.e., render judgments about quality). We propose using the more accurate term *student ratings of instruction* (SRIs), since "ratings" refer to data that require interpretation. Considering student ratings as data rather than evaluations puts them in a proper perspective (Benton & Cashin, 2014).

Wieman's first criticism against SRIs is that students cannot judge the effectiveness of an instructional practice unless they compare it to ones they have already encountered. If all students have experienced is lectures, he argues, how can they judge the quality of other methods compared to lectures? In our view, it depends on the questions we ask students. Questions should focus on teacher behaviors that have been shown to correlate with student progress and that cut across multiple teaching approaches. Examples include displaying a personal interest in students, stimulating students' intellectual efforts, and inspiring students to set and achieve goals that really challenge them. Such behaviors are connected to Chickering and Gamson's (1987) principles of good practice and can be demonstrated to students whether instructors employ lecture, discussion, or other active learning strategies.

According to Wieman, SRIs are invalid because people are poor at evaluating their own learning. Again, it may

depend on how the question is designed. We ask students to describe the amount of progress they have made on 12 learning objectives, only some of which may be relevant to the current course. They respond using a 5-point scale, ranging from *No apparent progress* to *Exceptional progress*. In this way we are able to distinguish between self-reported progress on objectives emphasized in the course from those on objectives not emphasized. If students cannot assess their learning, as Wieman claims, their ratings of progress on various objectives would not be consistent with the extent to which such objectives are emphasized in the course. Nevertheless, our analyses reveal that students report greater progress when the instructor indicates the objective was *Essential* compared to *Important*, and they report greater progress on *Essential* and *Important* (i.e., relevant) objectives than on those identified as being of *No or minor importance*.

Moreover, the correlations between student self-reported learning and actual achievement are comparable to those between *teachers'* judgments of students' achievement and students' *actual* achievement (see Sudkamp, Kaiser, & Moller, 2012, for a meta-analysis of 75 studies). Although the studies reviewed by Sudkamp et al. were conducted on secondary teachers, one would expect most college and university instructors—who typically have larger class sizes, less frequent contact with students, fewer opportunities to observe students' performance, and less educational preparation than secondary teachers in how to create valid and reliable assessments—would make even poorer predictions of students' actual achievement. We, therefore, postulate students' ratings of how much they have learned would be no worse than the judgments by their instructors.

Wieman goes on to cite Clayson (2009), who reviewed 17 articles across 42 datasets, containing 1,115 course sections, as evidence that SRIs fail to meet his first criterion of correlating with the desired student educational outcome. It is true that Clayson found, in general, a small positive association between measures of learning and SRIs. However, Clayson reported a raw average correlation of .33 and a median of .41, which is similar to what Cohen (1981) and Feldman (1993) found in their reviews a few decades earlier. So, contrary to what Wieman contends the magnitude of the correlations have *not* decreased in recent years, and they still do reveal a positive relationship between ratings and students' actual achievement. Clayson (2009) also states the relationship between learning and SRIs is valid "to the extent that the student's perception of learning is valid" (p. 27). We agree. We also contend that the relationship is valid to the extent that the instructor's measure of learning is valid. This may be why Clayson (2009) found higher correlations reported in studies published in education/psychology journals. Unlike their colleagues in other academic disciplines, those in education and psychology generally have taken coursework in psychometrics, which usually cover how to construct valid and reliable tests and writing assignments. When the tool used to measure learning exhibits adequate reliability, the correlation between learning and SRIs tends to be high, other things being equal.

Wieman claims another shortcoming of SRIs is revealed when student ratings are the same before and after instructors transform their teaching methods. Without knowing detailed design of the observation, we find it impossible to evaluate the legitimacy of his anecdotal statement. However, we believe an alternative explanation may account for his observation. Instructors sometimes fail to explain to students their reasoning behind employing a certain teaching method or emphasizing a certain type of learning (i.e., active learning). Communication about purposes is essential.

Next, Wieman argues that the correlation between SRIs and the desirable student outcomes must hold over a broad range of contexts and courses, and be much larger than the correlations with other factors beyond the instructor's control. While Wieman claims student ratings "fall far short of meeting" the criterion (p. 9), research has demonstrated the opposite. In Cohen's (1981) meta-analysis of data from 67 multi-section courses across 40 studies, he consistently found student final exam scores correlated positively with self -ratings of learning. To be included in Cohen's analysis, the study had to provide data from actual college classes rather than experiments. In a study involving over 50,000 classes across multiple disciplines, Centra (2003) examined the relationship between the grade students expected to receive in a course and their ratings of the quality of instruction. Controlling for class size, teaching method, and student ratings of progress on learning outcomes, expected grade generally had no effect on ratings of the instruction. However, student self-ratings of their learning consistently did.

We are especially disappointed that in Wieman's example to illustrate factors that may bias student ratings (p. 9), he includes attractiveness and gender of instructors in addition to class size, subject matter, and student level. Although a physical attractiveness stereotype does exist, its biasing effects vary based on the particular inferences individuals make. In their meta-analytic review of 76 studies, Eagly, Ashmore, Makhijani, and Longo (1991) found physical attractiveness had the strongest impact on perceived

social competence, followed by potency, adjustment, and intellectual competence, and had almost no influence on integrity and concern for others. While physical appearance may play a role in affecting student ratings (although a search using "student ratings of instruction" or "student ratings of teaching" and "physical attractiveness" found no hits), it is not unreasonable to expect its biasing effects are fairly limited, given student ratings instruments do *not* focus on instructors' social competence.

In the most comprehensive study of gender and student ratings to date, Centra and Gaubatz (2000) analyzed actual student ratings data (rather than data from simulations) across a large number of two- and four-year institutions, involving a variety of academic disciplines. They found only a small student-gender by teacher-gender interaction, particularly female students' preference for female instructors. Although the effects were statistically significant, they were moderate and would most likely not affect personnel decisions. The authors speculated that the higher ratings female instructors received from female students, and sometimes from male students, might have been a reflection of student preferences for certain teaching styles. Women in their study were more likely than men to use discussion than lecture, and they were perceived as more nurturing to students. Others (Feldman, 1993; Schulze & Tomal, 2006) also found no evidence of meaningful gender bias in student ratings.

Finally, Wieman argues student ratings provide little guidance for improvement because correlations are high among the items. Actually, student ratings on IDEA's 20 teaching methods are differentially related to ratings of student progress on relevant objectives. Employing Bayesian Model Averaging on data from nearly 500,000 classes across a 10-year period, Benton et al. (2015) found unique models, which varied somewhat by class size, for each of 12 learning objectives. For example, explaining the reasons for criticisms of student' academic performance is critical for developing creative capacities, appreciating intellectual/cultural activity, and expressing oneself orally or in writing. However, it is less important in other types of learning. Similarly, stimulating students to intellectual effort is strongly associated with progress on gaining factual knowledge and learning fundamental principles but less so with learning to apply course material.

IMPLEMENTING A COMPREHENSIVE TEACHING EVALUATION SYSTEM

We agree that institutions need a system for evaluating teaching effectiveness that incorporates multiple measures, such as quality of course design (especially for online courses), student products (e.g., creations, projects, papers), ratings by trained peers, teaching portfolios, and so forth. Hoyt and Pallett (1999) convincingly made this case in *IDEA Paper No. 36, Appraising Teaching Effectiveness: Beyond Student Ratings*. Student ratings are only one source of data; they must be combined with additional evidence so that administrators can make an informed judgment about teaching quality. Nonetheless, students' voices must continue to be heard. They take on tens of thousands of dollars of debt across their educational experience. We owe students at least one opportunity—if not more—during a semester to provide input about their learning experiences.

When assembling multiple information sources to render judgments about teaching effectiveness and instructional improvement, faculty should be selective and strategic in their accumulation of evidence (Halonen, Dunn, McCarthy, & Baker, 2012). An important principle to keep in mind is that the source identified should have first-hand knowledge of the performance being evaluated (Arreola, 2007). Students, for example, have first-hand knowledge of what occurs in the classroom, but they are not qualified to render judgments about the instructor's qualifications. In the following paragraphs we describe example sources of evidence that provide other important first-hand knowledge of teaching effectiveness.

External recognition

Evidence of effective teaching can be found in recognition the instructor receives external to the classroom. Being nominated for a teaching award, being invited to write a chapter about teaching, and being invited to share course materials are examples that the instructor's teaching is exemplary. Just as peer reviewers can offer judgments about the scholarly credibility of one's research, outside experts who recognize the instructor's teaching skills provide a good source of reputable knowledge.

Peer evaluations

Peer evaluation is most effective when the observer is trained in how to make classroom observations, review course materials, and give helpful feedback. Peers might want to adopt a standard tool, such as the *Teacher Behavior Checklist* (Keeley, Smith, & Buskist, 2006), or rubrics with normed benchmarks (Gurung & Schwartz, 2009). One might invite an expert in pedagogy from outside the discipline, such as someone working in the campus faculty development office. Such a person could provide helpful hints about changes the instructor might make in the classroom.

Examples of effective teaching

Much evidence can be found within the course to show the instructor has promoted active learning. Assignments, planned classroom activities, online course documents, and student products are but a few examples. Connections between goals and objectives, assignments, activities, and assessments would indicate the instructor has prepared a well-designed course. The instructor actually has the best knowledge of how the course was designed and the intentions behind it. This is, we believe, where Wieman's *Teaching Practices Inventory* fits.

Use of embedded assessment

The instructor might also provide evidence that shows students have made progress on important or essential learning outcomes either by student self-report or their performance on exams, projects, papers, and so forth. By aligning student performance with specific learning outcomes the instructor can determine whether students are grasping the essential objectives of the course and whether sections of the course need improvement.

Participation in professional development activities

Participation in professional development can signal the instructor's desire to improve. Such evidence should be accompanied by self-reflections or demonstrations of how the experience led to improvements in the course or approaches to teaching.

Evidence of "hazardous duty"

Some teaching assignments might be perceived as less desirable than others, such as large undergraduate general education classes, courses with especially challenging-to-teach content, and taking on an extra section to accommodate increased enrollments or a sick colleague. Such institution-centered behavior should be rewarded.

INTERPRETING MULTIPLE SOURCES OF EVIDENCE

Before giving equal weight to all sources of evidence, one should recognize the shortcomings in each of them. Although sometimes SRIs seem to receive the most scrutiny, each requires its own precautions. With respect to student ratings, faculty and administrators must recognize what students are and are not qualified to judge. As stated previously they are qualified to report what they observe happening in class. They are also capable of rendering judgments about how much they perceive they learned in the course, how well the course was delivered, and their desire to take the course. However, they are not qualified to judge the instructor's level of expertise, content of the course, or appropriateness of course goals. Administrators should also bear in mind that student ratings are influenced by student motivation, work habits, and class size. This is why IDEA controls for such variables in its adjusted scores and recommends that SRIs count no more than 30-50% of the overall teaching evaluation.

One must also be cautious when attempting to interpret information provided by the instructor. Some people are not comfortable reflecting on their own behavior, and in some cultures instructors might find it inappropriate to speak highly of themselves. Others may not believe they share responsibility for whether or not students learn, believing instead the responsibility lies in the hands of the students (Zakrajsek, 2006). This makes fair self-assessment especially difficult.

For peer observations and peer review of course materials additional cautions must be raised. Faculty may differ in their philosophies and approaches to teaching. Suppose a senior faculty member takes a teacher-centered, lecture-based approach to instruction. Would he or she be objective if called upon to evaluate a junior faculty member who applies a student-centered approach with active learning (Zakrajsek, 2006)?

With respect to external recognition and participation in professional development, the evaluators must scrutinize their value and learn as much as they can about awards and nominations or invitations. What was the quality of the professional development the faculty member attended? Were its goals aligned with those of the institution or department? Was it truly connected to better teaching?

Finally, faculty should have input into what constitutes "hazardous duty." Some argue, for example, that online courses are more work for the instructor than teaching face to face. It probably depends on how often instructors participate in online discussions, how quickly they respond to e-mails, and whether feedback to students is prompt and meaningful.

PUTTING IT ALL TOGETHER

Arreola (2007) defined five broad skill dimensions of college teaching displayed in Table 1. These are cross-tabulated with four information sources. In the paragraphs that follow we describe each dimension and explain our rationale for the best source(s) for assessment.

Content expertise includes skills, knowledge, and competencies in the specific subject matter being taught. Peers in the specific content area and external experts are best qualified to judge this aspect of teaching. However, the instructor can provide evidence through written self-reflection about relevant knowledge and experience, which may include recent

professional development. Administrators who are experts in the content area can also be a potential information source.

Instructional design skills are enlisted in creating learning experiences that support student achievement of learning outcomes. Although students are not qualified to evaluate the quality of instructional design, they can report their observations about what occurred and their perceptions of what and how much they learned. Peers, instructional leaders, and administrators would be the best judges of instructional design skills. However, the instructor can be another source via self-reflections about the rationale behind various design elements. Again, Wieman's inventory fits nicely here.

Instructional delivery skills refer to human interaction skills that help to promote student learning and motivation, including communication and educational technology skills. Students are the best source of information because they are the ones who interact with the instructor during the course.

Instructional assessment skills are the abilities to design tools, procedures, and strategies for assessing student achievement and providing meaningful feedback. We believe all four above-mentioned sources can contribute valuable information. Instructors can provide evidence of the coherence between learning objectives, tests, and assignments. Students can give their impressions about whether the assessments appeared fair and whether the feedback was meaningful. Peers and administrators can offer an objective view about the quality of the assessment instruments.

Course management skills refer to how well the instructor manages resources and facilities to provide a supportive learning environment. Examples include ordering and managing laboratory equipment, coordinating guest lecturers, purchasing software, monitoring and updating the course Web site, getting grades in on time, and completing all course-related "paperwork." The administrator is the best source of information about these activities.

Table 1
Skill Dimensions and Information Sources of Assessment of Teaching Performance

Skill Dimensions	Information sources			
	Instructor	Students	Peers	Administrator
Content expertise	YES	NO	YES	MAYBE
Instructional design skills	YES	YES	YES	NO
Instructional delivery skills	NO	YES	NO	NO
Instructional assessment skills	YES	YES	YES	YES
Course management skills	NO	NO	NO	YES

CONCLUSION

Student ratings of instruction should always have a place in the assessment of teaching effectiveness. Student voices are critical, not only because they provide some quality control but also because students have first-hand experience of what actually occurs in the classroom. But administrators who rely solely on student input when making summative decisions about teaching or when helping instructors to improve are skirting their responsibilities. Multiple sources of information are necessary for cross verification and for overcoming potential biases in any single measure.

IDEA has provided leadership in its mission to improve learning in higher education through research, assessment, and professional development. The research behind its instruments offers solid evidence for the reliability and validity of interpretations for both summative and formative feedback. When combined with other sources of evidence we expect IDEA will offer helpful feedback to instructors in the years to come. But, as always, we remain committed to continually refining and updating our services so to better serve higher education.

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Arreolo, R. A. (2007). Developing a comprehensive faculty evaluation system (3rd ed.). Bolton, MA: Anker Publishing Company, Inc.

Benton, S. L., & Cashin, W. E. (2014). Student ratings of instruction in college and university courses. In Michael B. Paulsen (Ed.), *Higher Education: Handbook of Theory & Research*, Vol. 29 (pp. 279-326). Dordrecht, The Netherlands: Springer.

Benton, S. L., Duchon, D., & Pallett, W. H. (2013). Validity of self-reported student ratings of instruction*. Assessment & Evaluation in Higher Education, 38,* 377-389.

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). *IDEA Research Report #8: Validity and reliability of IDEA Teaching Essentials*. Manhattan, KS: The IDEA Center.

Benton, S. L., Li, D., Brown, R., Guo, M., & Sullivan, P. (2015). *IDEA Technical Report No. 18: Revising the IDEA Student Ratings of Instruction System*. Manhattan, KS: The IDEA Center.

Brinko, K. T. (1990). Instructional consultation with feedback in higher education. *Journal of Higher Education*, *61*, 65-83.

Centra, J.A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44,* 495-518.

Centra, J. A., & Gaubatz, N. B. (2000). Is there a gender bias in student evaluations of teaching? *Journal of Higher Education*, *70*, 17-33.

Chickering, A. W., & Gamson, Z. (1987). Seven principles of good practice in undergraduate education. *American Association for Higher Education Bulletin, 39*, 3-7.

Clayson, D. E. (2009). Student evaluation of teaching: Are they related to what students learn? *Journal of Marketing Education, 31*, 16-30.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*, 281-309.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but…: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, *110*(1), 109–128. doi:10.1037/0033-2909.110.1.109

Feldman, K. A. (1993). College students' views of male and female college teachers: Part II--Evidence from students' evaluations of their classroom teachers. *Research in Higher Education*, *34*, 151-211.

Gurung, R. A. R., & Schwartz, E. (2009). *Optimizing teaching and learning: Pedagogical research in practice*. Malden, MA: Blackwell.

Halonen, J. S., Dunn, D. S., McCarthy, M. A., & Baker, S. C. (2012). Are you really above average? Documenting your teaching effectiveness. In R. A. R. Gurung & B. M. Schwartz, (Ed.), *Evidence-based teaching for higher education*, pp. 131-150. Washington, D. C.: American Psychological Association.

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.

Hativa, N. (2013a). *Student ratings of instruction: A practical approach to designing, operating, and reporting*. Oron Publications.

Hativa, N. (2013b). *Student ratings of instruction: Recognizing effective teaching*. Oron Publications.

Hoyt, D. P. (1973). Measurement of instructional effectiveness. *Research in Higher Education, 1,* 367-378.

Hoyt, D. P., & Lee, E. (2002). *IDEA Technical Report No. 12: Basic data for the revised IDEA system.* Kansas State University, Manhattan, KS: The IDEA Center.

Hoyt, D. P., & Pallett, W. H. (1999). *IDEA Paper No. 36, Appraising teaching effectiveness: Beyond student ratings.* Manhattan, KS: The IDEA Center.

Keeley, J., Smith, D., & Buskist, W. (2006). The Teacher Behavior Checklist: Factor analysis of its utility for evaluation teaching. *Teaching of Psychology, 33*, 84-91.

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

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.

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*, *27*(5). (pp. 27-44). San Francisco: Jossey-Bass.

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

Schulze, E. & Tomal, A. (2006). The chilly classroom: Beyond gender. *College Teaching*, *54*(3), 263-270.

Sudkamp, A., Kaiser, J., & Moller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*, 743-762.

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

Zakrajsek, T. (2006). Using evaluation data to improve teaching effectiveness. In P. Selden, Evaluating faculty performance: A practical guide to assessing teaching, research, and service (pp. 166-180). Bolton, MA: Anker Publishing Company, Inc.

T: 800.255.2757
T: 785.320.2400

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