IDEA Student Ratings of Instruction and RSVP

IDEA Paper #66 • September 2017

Stephen L. Benton and Dan Li • The IDEA Center

Abstract

Principles of effective assessment, abbreviated as RSVP, are reviewed within the context of the IDEA student ratings of instruction (SRI) system. *Reliability*, or consistency in scores, is high for IDEA SRI at the class and instructor levels. Internal consistency is also high for all factor subscales. *Standardization* of administration and scoring enables IDEA users to compare their scores with other classes in the overall research database, the respective academic discipline, and institution. Evidence for *validity* is found in (a) correlations between IDEA SRI and other variables, (b) multidimensional internal structure (meaning SRI measure two or more teacher qualities or constructs), (c) beneficial consequences of ratings, (d) logical item development, and (e) analysis of response processes. The *practicality* of IDEA SRI comes from its ease of administration and interpretation and its many helpful resources. Taken together, RSVP evidence supports using IDEA SRI as one, *but not the only*, source of evidence in formative, summative, and programmatic decisions that consider teaching effectiveness.

Keywords: student ratings of instruction, reliability, standardization, validity, practicality

"Many of the rating scales developed by faculty committees in colleges and universities do not meet even the most basic criteria for psychometric quality required by professional and legal standards"
(Berk, 2013, p. 34)

At first glance, the reader may wonder what IDEA student ratings of instruction (SRI) have to do with "RSVP" (from the French phrase *répondez s'il vous plaît*, which translates to "please reply"). Nothing. However, IDEA attempts to align its SRI system with several principles of effective assessment—*reliability* (R), *standardization* (S), *validity* (V), and *practicality* (P) (Ormrod, 2014). This paper will explain what these concepts mean, why they are important in assessment, and the extent to which IDEA SRI adheres to them.

When looking at an SRI class report, instructors may at times question whether students responded consistently or haphazardly. Did they take the ratings seriously? Was the SRI instrument a good assessment of what students learned and the behaviors the instructor exhibited in the classroom? Was the knowledge gained worth the time invested? These are legitimate questions, because ratings affect decisions that administrators make and modifications that instructors make to their courses. Therefore, instructors must make sound professional decisions about instrument selection or construction, administration, analysis, score reporting, and interpretation. Whether developing a home-grown SRI instrument or obtaining one produced by a national publisher, such as IDEA, the first task is to determine whether it follows important principles of good assessment.

Reliability

Reliability refers to the consistency in scores across repeated instances of administering an assessment instrument (AERA, APA, NCME, 2014). If faculty are to trust a student ratings system, they must be confident that the results will be similar for their class whether students complete the ratings on a Monday or a Wednesday, or regardless of the students' current mood. Reliability is especially important when the consequences of decisions based on the SRI have lasting impact. If the SRI determines decisions that cannot be easily reversed, such as continued employment, tenure, promotion, and merit pay, a high degree of reliability is necessary. If, on the other hand, student ratings are part of a balanced evaluation system considered along with other information sources (e.g., peer/administrator ratings, self-reflection), or if an erroneous initial decision can be reversed, a moderate level of reliability may be acceptable.

Multiple factors can affect reliability. First, longer surveys tend to yield higher reliability than shorter ones. As length

increases with additional items, reliability, in general, increases (AERA et al., 2014). This is because results that are based on multiple measurements tend to be more precise (on average) than results based on few measurements. Second, student preparedness can impact reliability. All students, for example, should be given the same set of instructions on how to complete the instrument. Otherwise, the difference in instructions introduces irrelevant variability in the responses. Third, variations in the physical environment can have an influence. Ideally, all students would fill out the form at the same time in the same format (e.g., on a mobile device during class). Such standardization is more difficult in Web-based administrations. Because of these and other factors, the track record for the reliability of home-grown instruments varies considerably (Berk, 2013).

Measurement Error

Reliability tends to be high for IDEA SRI most likely because (a) the instrument is of sufficient length, (b) online-administration procedures are clear and consistent, (c) items are clearly written at a reading level comprehensible to most students, (d) instructors generally exhibit the same kinds of teaching behaviors from one day to the next, and (e) students are typically consistent in their perceptions. Nonetheless, in practice, some inconsistency, or measurement error, occurs. The notion of error rests on true-score theory: A person's observed score (X) is equal to his or her average or true score (T) plus and minus the error score (E). The amount of error (E) in scores can only be estimated and is typically represented by the *standard error of measurement (SEM)*. The smaller the *SEM*, the better the SRI's reliability. *SEM* is, then, the flip side of reliability. As one goes up, the other goes down. *SEM* can be used to form a *confidence band*, or a range of values, around an instructor's observed SRI class mean score. The confidence band indicates the probability that an instructor's true score—or hypothetical average score—could fall within a range of two values. The greater the test score's reliability, the smaller the range. Perfect reliability is reflected in zero *SEM*.

SEM is, therefore, an indication of the inconsistency in measures of the same thing or, in the case of student ratings, in classmate perceptions of the same instructor. Unpredictable fluctuations in scores, or *random errors*, can come from two primary sources—either from within the raters themselves or from external factors. Internal sources of error arise from instabilities in student motivation, attention, and recall of what occurred in the class. Variations in the conditions surrounding the administration of the SRI (e.g., delivery format, environmental distractions, website difficulties) contribute to external sources of error. These random errors do not have a constant size or direction. They may occur for some respondents and not for others, which is why they are considered random.

When class size is at least 15 students, *SEM* is less than 0.3 on the 5-point scale for most IDEA items (see Table 1). For the items "Overall, I rate this instructor an excellent teacher" (excellence of teacher) and "Overall, I rate this course as excellent" (excellence of course), *SEM* is approximately 0.25. For example, if an instructor achieves a mean rating of 4.0 on either of those two items, the 68% confidence interval would extend from 3.75 to 4.25.[1] SEM is thus a reminder that every SRI score contains some error. In an effective teaching-evaluation system, faculty and administrators realize that any measure they use to assess perceptions of teaching effectiveness—be it SRI, peer observations, or self-ratings—is subject to random errors.

In contrast to random errors, *systematic errors* are inaccuracies that affect ratings in a consistent rather than an unpredictable manner. Examples of systematic errors would be differences between ratings of students in the same class who complete ratings on paper versus online, and differences between students who complete the SRI on a mobile device versus a computer. Such systematic errors generally do not contribute to *SEM*. However, they contribute to construct irrelevance, which reduces validity but not reliability.

Class-Level Reliability

When evaluating the reliability of an instrument, the first question to consider is whether the ratings by students within the same class are consistent. Consistency at the class level enables instructors to make interpretations that generalize across most students. If students within the same class vary substantially in their overall ratings of the course or instructor, perceptions depend on the individual student. Student feedback is thus less helpful in deriving general impressions of teaching, the course, and changes that should be made.

Class-level reliability can be measured by computing the *within-group interrater reliability coefficient* (James, Demaree, & Wolf, 1984). Coefficients range between .00 and 1.00; the higher the value, the more consistent students are in their perceptions of what they experienced in the class. Class-level reliability is high for IDEA SRI. Li and colleagues (Li, Benton, Brown, Sullivan, & Ryalls, 2016), for example, analyzed 2,426 classes, ranging in size from 15 to 34 students (average class size = 23). Reliability coefficients for IDEA SRI individual items ranged from .67 to .96, with all but one at or above .77 (see Table 1). Such high coefficients indicate students within the same class tend to perceive their experiences in the same way.

[1] The 90% confidence interval would range from 3.6 to 4.4.

Table 1 • *Within-Group Interrater Reliability Coefficients and Standard Errors of Measurement of Items on the Diagnostic Feedback Instrument in Medium-Size Classes (15–34 students)*

Item	*M*	*SD*	$r_{wg}(I)$	*SEM*
1. Helped students answer own questions	4.16	0.53	.86	0.20
2. Helped interpret subject matter	4.10	0.61	.82	0.26
3. Encouraged self-reflection	4.31	0.50	.88	0.17
4. Demonstrated significance	4.37	0.47	.89	0.16
5. Formed teams	3.89	0.81	.67	0.47
6. Made clear how topic fits	4.31	0.50	.88	0.18
7. Provided meaningful feedback	4.13	0.60	.82	0.25
8. Stimulated intellectual effort	4.11	0.53	.86	0.20
9. Encouraged using multiple resources	4.09	0.56	.84	0.22
10. Explained clearly	4.20	0.61	.82	0.26
11. Related to real life	4.33	0.52	.87	0.19
12. Created service opportunities	4.04	0.58	.83	0.24
13. Introduced stimulating ideas	4.18	0.56	.84	0.22
14. Involved in hands-on	4.01	0.67	.78	0.31
15. Inspired ambitious goals	4.00	0.58	.83	0.24
16. Asked diverse students to share ideas	4.01	0.68	.77	0.33
17. Asked students to help others	4.07	0.57	.84	0.23
18. Required originality	4.18	0.52	.86	0.19
19. Encouraged out-of-class contact	4.08	0.56	.84	0.23
20. Understanding subject matter	4.11	0.47	.89	0.16
21. Diverse perspectives	3.81	0.63	.80	0.28
22. Applications	4.07	0.50	.87	0.18
23. Professional skills and viewpoints	4.04	0.51	.87	0.18
24. Team skills	3.71	0.66	.78	0.31
25. Creative capacities	3.58	0.65	.79	0.30
26. Broad liberal education	3.72	0.60	.82	0.25
27. Communication skills	3.75	0.64	.79	0.29
28. Information literacy	3.89	0.55	.85	0.21
29. Ethical reasoning	3.76	0.64	.80	0.29
30. Critical analysis	3.92	0.57	.84	0.23
31. Civic engagement	3.87	0.60	.82	0.25
32. Quantitative literacy	3.70	0.62	.81	0.27
33. Amount of coursework	3.34	0.49	.88	0.17
34. Difficulty of subject matter	3.35	0.52	.87	0.19
35. Usually more effort on academic work	3.83	0.30	.96	0.06
36. Wanted course regardless of instructor	3.65	0.50	.88	0.18
37. Self-efficacy	3.90	0.37	.93	0.10
38. Background preparation	3.72	0.46	.89	0.15
39. Excellent instructor	4.28	0.60	.82	0.25
40. Excellent course	4.11	0.58	.83	0.24

Note. N = 2,426. Reprinted by permission of The IDEA Center.

Instructor-Level Reliability

Class-level reliability is a necessary condition for *instructor-level reliability*, which is the consistency in ratings of the same instructor across multiple classes. Instructor-level reliability is essential for confidence in inferences about how well an instructor teaches, in general, which are made as administrators consider merit pay increases, retention, promotion, and tenure. The *interclass-reliability coefficient*, which indicates the extent to which the SRI vary among different instructors and exhibit similarity regarding the same instructor (Gillmore, 2000), is the typical measure of instructor-level reliability. This can be distinguished from the *intraclass-correlation coefficient*, described in the next section as *Cronbach's alpha*, which is used when item responses are organized into groups. A high interclass coefficient indicates that differences in ratings among distinct instructors are greater than differences across classes taught by the same instructor. Again, values range from .00 to 1.00, where 1.0 indicates perfect reliability.

Instructor-level reliability is high for IDEA SRI, based on individual item interclass coefficients calculated among 2,500 instructors who had each been rated in at least five classes (Benton, Li, Brown, Guo, & Sullivan, 2015). Figure 1 shows the plots for IDEA's two overall summary measures: "Overall, I rate this instructor an excellent teacher" and "Overall, I rate this course as excellent." When at least three classes have been rated, reliability coefficients are above .80. Coefficients for all items are .90 or greater when ratings have been conducted in at least six classes. Consequently, IDEA has for years recommended that ratings be collected in at least six classes, preferably eight, before summative decisions are made about an instructor.

Internal Consistency

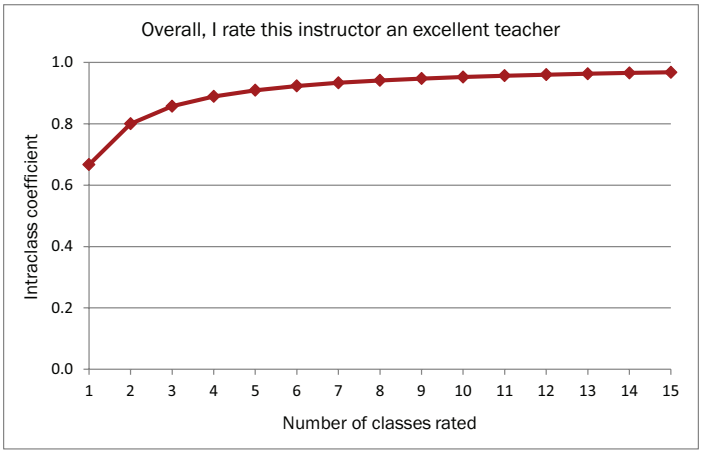
Up to this point, I have focused on the reliability of student responses to individual items. However, often student-ratings instruments contain subscales, which are collections of items measuring a common construct. *Internal consistency reliability*, an estimate of how consistently the items within a scale measure the same construct, is typically measured with *Cronbach's alpha* (α) (Cronbach, 1951) or the *intraclass coefficient*. Cronbach's α is a ratio of (a) the number of items in the subscale squared and then multiplied by the average covariance between items and (b) the sum of all subscale item variances and covariances. Values typically range from .00 to 1.00, with higher coefficients indicative of greater internal consistency. Although IDEA does not provide scores for its subscales, which are described in the section on validity, the intraclass coefficients are all above .90 (Li et al., 2016), which indicates high reliability.

Standardization

Whereas *reliability* refers to consistency in ratings, *standardization* concerns consistency in directions, administration, physical conditions, and scoring for all students completing the SRI. IDEA bolsters standardization by providing explicit instructions on getting started, administering the survey, and using the feedback from students appropriately. Instructors should first read the information provided on the website and review the survey a few days before administering it. Next, they should set a time frame—start and end dates—for when the survey will be available to students. If either IDEA's Diagnostic Feedback or Learning Essentials is administered, instructors should select the relevant objectives ("essential" or "important") for their course before administering the survey. The Choosing Learning Objectives video is a helpful guide, and, usually, no other special training is required. The selected objectives should be consistent with the ones emphasized in the course.

Standardization also entails orienting students to procedures for completing the ratings. The instructor can prepare students by demonstrating how to access the survey and showing them a sample. The instructor can motivate students by explaining the important decisions that administrators may make based, in part, on their responses. Other suggestions for encouraging students to complete the ratings can be

Figure 1 • Plots of Inter-class Reliability Coefficients for Overall Summary Measures



Note. Reprinted by permission of The IDEA Center.

found at http://www.ideaedu.org/Resources-Events/Support-For-Current-Clients/Best-Practices-for-Increased-Response-Rates.

Finally, standardization also necessitates protecting the integrity of student-ratings scores. IDEA SRI, powered by Campus Labs, provides state-of-the-art data security. Confidentiality is protected, and neither students nor instructors can alter the ratings once they have been submitted.

Standardization is necessary for reducing random error and developing norms or, in the case of student ratings, comparative scores under standard conditions of administration and scoring. IDEA offers comparative scores so that instructors can benchmark their own against all classes in its database, classes in the most relevant academic discipline, and classes in the respective institution. *Standard T-scores* express an instructor's raw score average on the 5-point scale in standard deviation units from the comparative group mean. A score of 50 represents the mean in the respective comparison group (e.g., academic discipline), and a difference of 10 above or below the mean is equal to one standard deviation.

Validity

Whereas reliability and standardization concern consistency in measurement, administration, and scoring, *validity* addresses whether the interpretation from the instrument is appropriate and whether it is being used for its intended purpose. Validity is, therefore, the most important quality to consider when either developing or choosing an SRI (AERA et al., 2014). No one can say that any given measure is completely valid, any more than anyone can credibly claim to be completely honest. In each case, one can only provide evidence to support the assertions of validity and honesty. Validity, then, pertains to ratings and inferences, not the measure itself.

How Is Reliability Related to Validity?

Reliability creates limits on validity. For example, what if the observations of how frequently an instructor used each of IDEA's 19 teaching methods differed greatly from one student to the next? The instructor would then have to question whether the SRI items were functioning as designed, whether students were paying attention in class, or whether they were taking the ratings seriously. The inconsistency in measurements would threaten the accuracy of interpretations made from the ratings.

However, although reliability is important, it does not guarantee validity. For example, students would most likely be consistent (i.e., reliable) in their responses to the following scale, but it would fall short of being a valid measure of teaching effectiveness:

1. Did the instructor greet students at the beginning of each class?
2. Did the instructor wear glasses?
3. Did the instructor provide a syllabus?

Reliability, therefore, is a necessary, but not a sufficient, condition for validity.

Construct Underrepresentation

Construct underrepresentation is the degree to which an instrument fails to measure important aspects of the trait being assessed (AERA et al., 2014). In the case of SRI, there are important aspects of teaching effectiveness that students are unqualified to judge and that ratings, therefore, underrepresent (Benton & Cashin, 2014). Examples of teaching qualities not fully captured and evaluated by SRI include the instructor's subject-matter knowledge and commitment to teaching; the quality of course design; the appropriateness of goals, learning objectives, and course content; and the quality of tests, among other course elements. As Wiggins (1998, *p*. 248) cautions, "a single score hides more than it reveals." Consequently, student ratings should never serve as the only measure of teaching effectiveness.

Construct Irrelevance

Another word of caution concerns the possibility that processes or factors extraneous to its intended purpose could affect an SRI score, thereby creating *construct irrelevance* (AERA et al., 2014). Faculty and administrators are sometimes apprehensive, for example, about possible biases in student ratings. Numerous studies have examined whether ratings are affected by the instructor's race and gender, student interest and motivation, and a student's expected grade, among other variables (for a review, see Benton & Cashin, 2011). On the whole, though, well-designed student-ratings instruments are relatively unaffected by a variety of potential biases (Benton & Ryalls, 2016; Marsh, 2007, *p*. 372).

Before concluding that an instrument is biased, one should consider Marsh's (2007) word of caution: "Bias exists when a student, teacher, or course characteristic affects the evaluations made either positively or negatively *but is unrelated to any criteria of good teaching*" (*p*. 350). In that sense, correlations between student ratings and extraneous variables, such as class size, student interest and motivation, and expected grade are not necessarily biases. Students in small classes and students who are more interested in the course *do* tend to learn more, earn higher grades and, consequently, assign their instructor higher-than-average ratings. In such cases, high ratings may reflect good teaching. Nonetheless, IDEA tries to control for extraneous influences by adjusting average course ratings for (a) class size; (b) student self-reported motivation, work habits, and background preparation; and (c) perceived difficulty of the subject matter (after removing the influence of the instructor) (Li et al., 2016).

However, what about instructor gender? Much attention has been directed toward studies that have explored possible

instructor-gender bias in student ratings (e.g., Flaherty, 2016). Nonetheless, instructor gender has no practically meaningful effects on IDEA SRI in either STEM (science, technology, engineering, mathematics) or non-STEM fields (Li & Benton, 2017). Average student ratings are very similar in courses taught by men and those taught by women, regardless of academic-discipline group. Moreover, the interaction effects of instructor gender and academic-discipline group (i.e., STEM vs. non-STEM) do not exert much influence. The most telling difference is observed not between men and women but between STEM and non-STEM instructors. (Notably, IDEA has for years provided separate comparative standard T-scores for the instructor's respective academic discipline.) Still, IDEA users may want to investigate whether gender differences in ratings exist on their campuses. At local levels, some differences could be meaningful, particularly if ratings are the only measure used in making summative decisions about teaching effectiveness.

Joint Responsibility of the SRI Developer and User

With all these issues surrounding validity, the SRI publisher or the faculty committee that develops a local instrument must gather supportive evidence of validity. In turn, the SRI user shares responsibility for proper use and interpretation of the instrument. Publishers should communicate how SRI scores should be interpreted; users should either follow those guidelines or provide evidence to support interpretations that differ from the ones recommended. For example, in its publications, IDEA recommends that student ratings should never be the sole measure of teaching effectiveness (Benton & Cashin, 2011; Benton & Ryalls, 2016; Ryalls, Benton, Barr, & Li, 2016). Thus, if institutions rely on IDEA SRI exclusively for summative evaluations of teaching, they are diverting from the instrument's intended purpose and should, therefore, provide evidence to support such a practice.

Evidence of IDEA SRI Validity

Now let's examine the evidence that supports the use of IDEA SRI for formative and summative evaluation. There are multiple ways to show evidence that either supports or refutes validity claims. As with partying, "the more, the merrier." To put it another way, multiple sources of relevant evidence in support of claims of validity bring greater confidence. Validity evidence for IDEA SRI comes from several sources, examples of which follow.

1. *Correlations with other conceptually relevant variables*. Relation to other variables is commonly regarded as *construct validity*, which is found in relationships between student ratings and another criterion assessed at the same time (e.g., correlations between student ratings of progress on learning objectives and the instructor's ratings of the same objectives for relevance to the course).
2. *Test-criterion relationships* demonstrate the relationship between student ratings and a relevant performance measure or criterion (e.g., SRI and exam performance).
3. *Internal structure* indicates the degree to which relationships among responses to items conform to an intended structure (e.g., intended subscales).
4. *Beneficial consequences of ratings* (e.g., improvements in the quality of teaching, improvements in student achievement).
5. *Logic of test content* is the connection between the items on a ratings scale and what the scale intends to measure (e.g., IDEA SRI contains learning objectives because it intends to measure student self-perceptions of progress).
6. *Student response processes that confirm item meaningfulness*. For example, interviews can be conducted with students to ascertain whether or not they interpret the items as intended.

1. *Evidence that student ratings correlate with other conceptually relevant variables*. In the IDEA SRI, both the instructor and students rate the same 13 learning objectives. Although they are each rating different entities—students rate themselves and instructors rate the objectives—their ratings are conceptually related. The instructor rates the objectives for relevance to the course, using a 3-point scale of *essential, important, or minor or no importance*. Students judge their progress on the objectives, using a 5-point scale, which ranges from *no apparent progress* to *exceptional progress*. Because they are assigning ratings based on experiences in the same course, the instructor's and students' scores are linked and therefore lend themselves to correlational analysis.

   IDEA correlates the instructor's rating on each objective with that of the average class rating across thousands of classes. The highest correlations between instructor and student ratings are found in corresponding objectives, which is evidence for *convergent validity*—relationships between two measures intended to assess the same or similar constructs. Conversely, lower correlations tend to occur in noncorresponding objectives, which supports the instrument's *divergent validity*—relationships between measures of different constructs (Benton et al., 2015; Li et al., 2016). In other words, students tend to report the greatest progress on objectives stressed by their instructor.

   Another example of construct validity evidence is found in the relationships between average student ratings of their own motivation, work habits, and background preparation and the instructor's overall ratings of the class as a whole on those same student characteristics. In a study by Benton and Li (2017), students responded to three items, using a 5-point scale (*definitely false* to *definitely true*): "As a rule, I put forth more effort than other students on academic work" (work habits), "I really wanted to take this course regardless of who taught it" (course motivation), and "My background prepared me well for this course's requirements" (background preparation), which are the three most important variables in IDEA's adjusted scores. Instructors, in turn, rated three characteristics of the class as whole: "student enthusiasm for the course," "student effort to learn," and "adequacy of students' background and

preparation for the course." Each instructor responded by indicating whether the circumstance "had a positive impact on learning," "had neither a positive nor a negative impact," or "had a negative impact on learning." "Cannot judge" was also an option.

Benton and Li (2017) conducted independent *t* tests to examine whether average student self-ratings of work habits, motivation, and background preparation were higher in classes where the instructor indicated the respective characteristic had a positive impact on learning, compared to those in which he or she reported a negative impact. Cohen's *d* (1988) was employed as a measure of effect size, where *small* = .20, *medium* = .50, and *large* = .80. Student self-ratings of motivation were higher in classes where instructors said student enthusiasm had a positive impact than in classes where it was reported to have a negative impact (Cohen's *d* = .75). In classes where instructors reported a positive impact of student effort, student ratings of work habits were higher than in classes where instructors perceived that it had a negative impact (*d* = .29). Finally, student self-ratings of background preparation were higher in classes where the instructor reported the adequacy of student background and preparation had a positive impact (*d* = .58).

In a similar study, Benton et al. (2015) found a connection between instructor perceptions of various course and student circumstances (e.g., physical facilities; desire to teach the course; students' levels of preparation, enthusiasm, and effort) and student ratings of progress on relevant objectives (PRO) and overall summary measures. For the most part, instructors who held positive views of course circumstances also had higher student ratings of PRO, excellence of the teacher, and excellence of the course than those who had negative perceptions. The strongest effects were consistently shown in student overall ratings of the excellence of the course (Benton et al., 2015). Instructor perceptions of course circumstances coincided, then, with students' impressions of the instructor, the course, and their overall learning.

Additional evidence of construct validity is shown in the connection between student ratings of progress on relevant objectives and the instructor's course requirements. For instance, student self-reported progress on communication skills is greater in courses where instructors emphasize writing skills than in courses where they do not. Moreover, the same pattern is seen in student ratings of progress on creative capacities, team skills, and critical thinking (Benton et al., 2015). In general, instructors that emphasize these skills have students who report greater progress on them.

Finally, IDEA SRI correlate positively with external student ratings of learning and teacher behaviors. For example, student ratings of progress on relevant objectives concerning factual knowledge, principles and theories, applications, professional skills and viewpoints, and interest in learning are highly correlated with external student ratings of teacher feedback and with congruity between learning outcomes and course activities (McAlpine, Oviedo, & Emrick, 2008). Also, student ratings of instructor clarity and helpfulness, conducted on a public Web-based ratings system external to IDEA, are significantly and positively correlated with ratings of overall instructor and course excellence on the IDEA SRI (Sonntag, Bassett, & Snyder, 2009).

2. *Evidence of test-criterion relationships*. One example of a test-criterion relationship is whether IDEA's student ratings of progress on relevant objectives correlate with how much students have learned in the course. Benton, Duchon, and Pallett (2011) investigated this question in a study of multiple sections of the same course taught by the same instructor. Students rated their progress on all IDEA SRI learning objectives, including two their instructor had identified as relevant to the course. Progress ratings on relevant objectives were significantly and positively correlated with student performance on four out of five exams (Pearson *r* correlation coefficients ranging from .19 to .41) and total course points (*r* = .32), which is evidence for convergent validity. However, student progress ratings of irrelevant objectives were not related to exam performance (coefficients ranging from –.11 to .09) and total course points (*r* = –.07), which supports divergent validity. So, there is a relationship between student perceptions of how much they have learned and performance on teacher-made assessments of learning.

3. *Evidence based on internal structure*. Because effective teaching requires the performance of many behaviors, most well-designed student-ratings instruments conform to a multidimensional structure (Hativa, 2014), meaning they measure two or more teacher qualities or constructs (Ackerman, Gierl, & Walker, 2003). For example, the 13 learning objectives in the IDEA SRI system are intended to emphasize different academic skills and abilities. Because instructors are well qualified to differentiate the relevance of the objectives of their course, one would expect their ratings to be multidimensional, which is in fact the case. Instructor ratings of relevance break out into four underlying dimensions, each having high internal consistency reliability (all Cronbach's alphas > .94): General Life Skills (i.e., critical thinking, communication skills, ethical reasoning/decision making, diverse perspectives, information literacy, and civic engagement); Professional Skills (i.e., professional skills and viewpoints, team skills, and applications); Cultural/Creative Development (i.e., broad liberal education, creative capacities); and Course-Specific Skills (i.e., understanding subject matter, quantitative literacy) (Li et al., 2016).

Just as faculty can discriminate between each objective's relevance, so too can students distinguish the progress they make on each objective. Student ratings of progress fall along two reliable dimensions: General Life Skills (α = 96) (i.e., diverse perspectives, communication skills, creative capacities, broad liberal education, ethical reasoning, critical thinking, civic engagement, and team skills) and Course-Specific Skills (α = .94) (i.e., quantitative literacy, understanding subject matter, applications, professional skills and viewpoints, information literacy) (Li et al., 2016).

Multidimensionality is also found in student ratings of how frequently the instructor used each of the 19 teaching methods. Students distinguish between two broad categories of teaching behaviors, each having high reliability: Instructor-Centered (α = .98), having to do with actions instructors take (e.g., "explained material clearly and concisely," "made it clear how topics fit into the course," "introduced stimulating ideas"), and Learner-Centered (α = .94), which describes teacher actions that facilitate active student learning (e.g., "formed teams," "involved students in hands-on activities," "asked students to share their experiences") (Li et al., 2016).

Another aspect of internal structure unique to IDEA are the distinct relationships between student ratings of teaching methods and progress on relevant learning objectives. The methods most highly correlated with progress on each objective follow predictable patterns. For example, the teacher behaviors of "made it clear how each topic fit into the course" and "stimulated students to intellectual effort beyond that required by most courses" are strongly correlated with student reported progress on the cognitive learning objectives of "understanding subject matter" and "applying course material"; however, those teaching methods are less important for acquiring team skills and developing creative capacities. Team skills and creative capacities benefit more from teacher actions that "involved students in hands-on projects" and "inspired students to set and achieve goals which really challenged them." Other examples of logical relationships between teaching methods and learning objectives may be found in Li et al. (2016).

4. *Evidence based on the consequences of ratings*. Perhaps one of the greatest sources of controversy surrounding student ratings are the intended and unintended consequences of using them. Although faculty can derive benefits—such as improvements in the quality of teaching, improvements in student achievement, recognition, salary increases, and promotion—negative consequences may follow as well (Berk, 2006). For example, sometimes instructors mistakenly believe that they can receive higher ratings by lowering standards and expectations for students. The opposite is true: Analyses of ratings in approximately 500,000 classes across more than 300 institutions reveal that instructors are more likely to earn high ratings when their students report that the instructor challenged them and had high achievement standards (Benton, Guo, Li, & Gross, 2013). Nonetheless, such misconceptions persist. Another unintended negative consequence is the poor practice of making student ratings the only measure of teaching effectiveness.

   Since its birth, IDEA's Diagnostic Feedback is intended primarily to benefit instructors by helping them to strengthen teaching practices associated with student progress on relevant learning objectives (Hoyt & Cashin, 1977). Some evidence indicates that students report greater progress when instructors discuss their IDEA reports with a peer or consultant (Burbano, 1987). In classes where the instructor received diagnostic feedback plus consultation after midsemester administration of IDEA, students reported significant improvement on 6 of 10 learning objectives in end-of-course ratings. Cohen's *d* ranged from .32 to .50, which represents approximately a one-third to one-half standard deviation improvement. In classes where an instructor received diagnostic feedback only, students reported significant improvement on 4 of 10 learning objectives (Cohen's d ranged from .18 to .44). In classes where instructors received no midsemester feedback or consultation, students reported no significant improvement on any objectives.

5. *Evidence based on test content*. A less empirical source of evidence addresses the practical question of how the SRI items were developed. Evidence based on test content concerns the themes, wording, and format of the items on an SRI. All the IDEA items were developed after careful thought, reviews of relevant literature, and the input of experts in the field. For example, of the 13 learning objectives, four were developed in 1969 based on reviews of Bloom's (1956) and Krathwohl, Bloom, and Masia's (1964) taxonomies of educational objectives and Deshpande and Webb's (1968) factor analysis of objectives endorsed by faculty. One new objective was added in 1972 to accommodate the need for assessing creative capacities. Three were then added in 1998 to address team skills, information literacy, and critical thinking, based on a survey of users from 32 institutions.

   Of the 19 teaching methods, four were developed in 1969, based on factor analyses of several existing instruments and reviews by faculty who had won teaching awards. Three new items were added in 1972 and seven more in 1998, based on the aforementioned survey of users. Of the six student and course characteristics, one was created in 1972, and three were added in 1998, based on the survey of users.

   For the current instrument, five new learning objectives, five new teaching methods, and two new student and course characteristics were added in 2016 in response to recommendations from focus groups and expert

panels. All items retained from earlier versions were reviewed by two focus groups representing veteran users of IDEA who varied in gender and ethnicity. Focus-group participants, who included educators from both public and private institutions, made suggestions for retaining, revising, or dropping items. Interviews were also conducted with IDEA staff, several of whom had advanced graduate degrees and years of experience working in the field of student ratings.

New items were proposed after a comprehensive review of the literature on student ratings and teaching and learning. Two expert panels then reviewed all retained and proposed items. The panels comprised content-area faculty experts and experts in teaching and learning, technology, measurement, faculty development, faculty evaluation, higher-education administration, and institutional assessment. Feedback was also obtained from IDEA users and nonusers via an online questionnaire. After conducting a content analysis of 25 faculty responses to the open-ended survey questions, additional revisions were made (Benton et al., 2015).

6. *Evidence based on response processes*. A final source of validity evidence comes from interviews conducted with college students as they responded to the items on the IDEA SRI. To ascertain whether or not students interpreted the items as intended, volunteers were recruited from college classes and were asked to think aloud as they answered each question. Selection of students was stratified by gender, age, academic major, class level, and English proficiency. Each interview took between 30 and 45 minutes. Afterward, recorded notes were analyzed to discover consistent patterns in the ways students understood and responded to individual items. Minor revisions were then made to some items based on students' responses (Benton et al., 2015).

In summary, validity evidence for IDEA SRI comes from research that demonstrates correlations with other conceptually relevant variables, test-criterion relationships, multidimensional internal structure, beneficial consequences of ratings, logical test content, and student response processes that confirm item meaningfulness.

Practicality

Up to this point, we have presented evidence that supports valid and reliable interpretations of IDEA SRI under standard conditions. However, all that would be of little value to users if the instrument itself were impractical to use. Compared to other sources of evidence used in the evaluation of teaching, student ratings are probably the most practical in terms of investment of time, people, and resources. Instructor self-ratings and ratings by peers require that faculty devote substantial hours to reflecting and writing. Upon completion, the result represents the perceptions of a single person. In contrast, student ratings require very little instructor time, and the feedback received represents the viewpoints of multiple individuals. Ultimately, though, the practicality of any SRI concerns whether instructors can administer them easily with the least disruption to class, whether instructors can understand the student feedback summarized in the class report, and whether the system offers helpful resources.

Ease of Administration

An important practical consideration is the ease with which the instrument is administered. Ease of administration is especially important in light of research that indicates that response rates to Web-based surveys are lower than those to paper surveys (Avery, Bryant, & Mathios, 2006; Benton, Webster, Gross, & Pallet, 2010; Layne, DeCristoforo, & McGinty, 1999). If students become frustrated by a system that has a poorly designed interface, they may be less likely to complete the ratings.

IDEA SRI, powered by Campus Labs, is supported by simple and brief Web-based directions that make starting, administering, and completing the survey easy for faculty and students. Campus coordinator resources are also provided, as well as technical support as needed. The system is compatible with most course management systems in a user-friendly, interactive online platform, along with a mobile interface. Items are written at a reading level comprehensible to most students. Although survey length is understandably a concern, it should not be the primary reason for adoption or rejection. For example, the completion rates for IDEA's 40-item Diagnostic Feedback and its 18-item Learning Essentials forms are the same (99%). The vast majority of students who start the surveys finish them, regardless of length. Finally, because the time required to complete the ratings is minimal, instructors teaching on campus can ask students to complete the survey in class, which should result in higher response rates.

Ease of Interpretation

Another consideration is whether class reports are easy to interpret. If faculty find it difficult to access the reports or if the reports are verbally dense, instructors may be less likely to put in the necessary time to understand the results. On IDEA SRI, faculty receive individual class reports online and as a printable PDF. The interactive online reports are accompanied by meaningful charts and do not contain overly technical information or complex statistics. Summative and formative feedback, as well as quantitative and qualitative information, are provided. Helpful recommendations are also made regarding actions the instructor might take to improve student progress on relevant learning objections.

Helpful Resources

A final practical point is what additional resources the developer offers. In addition to individual class reports, institutions should be able to access data for any course evaluated in the system. IDEA users can, for example, access course ratings for their institution through the IDEA Data Portal, which enables programmatic access to institutional surveys and data. IDEA partners use the system to integrate data with on-campus archives, learning-management systems, and so forth. Another useful IDEA product is the

Unit Summary Report, which provides results for programs and colleges that wish to have an aggregated report of all the classes in a unit or curriculum. Other services include consultation for interpreting reports, webinars, and online resources documenting the research behind the instruments as well as white papers on teaching and learning.

Summary

IDEA has accumulated multiple sources of evidence to support appropriate use of its student-ratings instruments. Reliability coefficients are high for all items at both the class and instructor levels, and internal consistency is high for all factor subscales. Through its standardization of administration and scoring, IDEA enables instructors to compare their scores with those of other classes in the overall research database and their respective academic disciplines and institutions. Evidence for validity is found in (a) correlations between IDEA SRI and other variables, (b) multidimensional internal structure, (c) beneficial consequences of ratings, (d) logical item development, and (e) analysis of response processes. The practicality of IDEA SRI comes from its ease of administration and interpretation and its many helpful resources. Taken together, the RSVP evidence summarized in this paper supports the use of IDEA SRI as one, but not the only, source of evidence in formative, summative, and programmatic decisions about teaching effectiveness.

Steve Benton is Senior Research Officer at The IDEA Center where he leads a research team that designs and conducts reliability and validity studies for IDEA products. He is a Fellow in the American Psychological Association and American Educational Research Association, as well as an Emeritus Professor of Special Education, Counseling, and Student Affairs at Kansas State University where he served for 25 years. His current research focuses on effective use of student ratings of instruction for improving teaching and learning.

Dan Li is Research Associate at The IDEA Center. She holds a B.A. from Huazhong University of Science and Technology, an M.A. from Marquette University, and a Ph.D. in Media, Technology, and Society from Northwestern University. Her previous research examined the social effects of online technologies, digital inequality, and parental mediation of television viewing. Her current work focuses on student ratings of instruction in higher education.

References

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement, 22*, 37–51.

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

Avery, R. J., Bryant, W. K., & Mathios, A. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations. *The Journal of Economic Education, 37*, 21–37.

Benton, S. L., & Cashin, W. E. (2011). *Student ratings of teaching: A summary of research and literature*. IDEA Paper No. 50. Manhattan, KS: The IDEA Center.

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. (2011). Validity of self-report student ratings of instruction. *Assessment & Evaluation in Higher Education. 38*, 377–389.

Benton, S. L., Guo, M., Li, D., & Gross, A. (2013). *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. (2017). *IDEA Research Note #6: Validity of the IDEA Student Ratings of Instruction student characteristics items*. 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.

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, 2002–2008 data*. Manhattan, KS: The IDEA Center.

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

Berk, R. A. (2013). *Top 10 flashpoints in student ratings and the evaluation of teaching*. Sterling, VA: Stylus.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals: Cognitive Domain*. Longman.

Burbano, C. M. (1987). The effects of different forms of student ratings feedback on subsequent student ratings of part-time faculty (Doctoral dissertation). University of Florida, Gainesville, Florida.

Cronbach, L. J. (1951). "Coefficient alpha and the internal structure of tests". *Psychometrika*, *16* (3): 297–334.

Deshpande, A. S., & Webb, S. C. (1968). Student perceptions of instructor teaching goals. III. *Internal structure of ratings (Research Memorandum 68-5)*. Atlanta: Office of Evaluation Studies, Georgia Institute of Technology.

Flaherty, C. (2016). Bias against female instructors. Inside Higher Education. https://www.insidehighered.com/news/2016/01/11/new-analysis-offers-more-evidence-against-student-evaluations-teaching

Gillmore, G. M. (2000). *Drawing inferences about instructors: the inter-class reliability of student ratings*. University of Washington, Seattle, WA: Office of Educational Assessment.

Hativa, N. (2014). *Student ratings of instruction: Recognizing effective teaching* (2nd ed.). Oron Publications.

Hoyt, D. P., & Cashin, W. E. (1977). *IDEA Technical Report No. 1: Development of the IDEA system*. Kansas State University, Manhattan, KS: The IDEA Center.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*, 85–98.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives, Handbook II: Affective Domain*. New York: David McKay Co., Inc.

Layne, B. H., DeCristoforo, J. R., & McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education, 40*(2), 221–232.

Li, D. & Benton, S. L. (2017) *IDEA Research Note # 6: Validity of the IDEA student ratings of instruction student characteristic items*. Manhattan, KS: The IDEA Center.

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.

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.

McAlpine, L. & Oviedo, G. B., & Emrick, A. (2008). Telling the second half of the story: Linking academic development to student experience of learning. *Assessment and Evaluation in Higher Education, 33*, 661–673.

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

Ryalls, K. R., Benton, S. L., Barr J., & Li, D. (2016). *Response to "Bias against female instructors."* Editorial Note. Manhattan, KS: The IDEA Center. http://ideaedu.org/research-and-papers/editorial-notes/response-to-bias-against-female-instructors/

Sonntag, M. E., Bassett, J. F., & Snyder, T. (2009). An empirical test of the validity of student evaluations of teaching made on RateMyProfessors.com. *Assessment and Evaluation in Higher Education, 34*, 499–504.

Wiggins, G. P. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco: Jossey-Bass.

T: 800.255.2757
T: 785.320.2400

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

Our research and publications, which benefit the higher education community, are supported by charitable contributions like yours. Please consider making a tax-deductible donation to IDEA to sustain our research now and into the future.