IDEA Research Report #8 • March 2015

Validity and Reliability of IDEA *Teaching Essentials*

Stephen L. Benton and Dan Li • The IDEA Center

Most institutions of higher education have some system in place for collecting student ratings of instruction (SRI) data from classes on a systematic basis. SRIs are a relatively simple way of gauging student perceptions of teaching and the course. Although sometimes besieged by controversy and misconceptions, student ratings can provide valid and reliable feedback to instructors (Benton & Cashin, 2011, 2014). After all, students are the ones personally affected by instruction, and they have substantial opportunities to observe actual teaching behaviors, which enhance evidence for validity. Moreover, student ratings can be viewed as the single most reliable measure of teaching effectiveness because they represent the observations of multiple raters across multiple occasions.

Since the initial development of IDEA SRI in 1975, much has changed about higher education. Many courses are offered entirely or partially online, most students and faculty have access to mobile devices, and an increasing percentage of instructors are employed part-time. Faculty and administrators need a student ratings instrument that is responsive to such changes and that offers quick and helpful feedback to both full- and part-time instructors. IDEA's *Teaching Essentials* was created, in part, to address those needs. The 12-item instrument measures student overall impressions of the instructor and course, observations of seven principal teaching methods practiced by the instructor, and three extraneous characteristics that can influence ratings. This report describes the validity and reliability evidence behind the development of *Teaching Essentials.*

VALIDITY

Validity concerns whether credible evidence exists to support appropriate use and interpretation of scores derived from some measure. In the context of student ratings, "appropriate use" rests in the hands of administrators and faculty who interpret ratings to make decisions about teaching effectiveness. Because student ratings are a necessary but insufficient source of evidence, IDEA recommends that they count no more than 30% to 50% of the overall teaching evaluation. Additional indicators of teaching quality should be considered (e.g., administrator and peer ratings, course materials, exams, student products, alumni ratings). In this section, we present evidence to support valid interpretations of data collected from TE.

Excellence of Instructor and Course as Overall Summary Measures

To develop an effective yet succinct instrument, we included two items to measure student overall impressions of the instructor and the course. As with IDEA's longstanding *Diagnostic Feedback* and *Learning Outcomes* instruments, students provide two summary judgments on the TE: *Overall, I rate this instructor an excellent teacher* and *Overall, I rate this course as excellent*.[1] They respond using a 5-point Likert scale (ranging from 1 = *Definitely False* to 5 = *Definitely True*). Evidence for convergent validity is found in the items' strong positive correlations with the amount of progress students report on relevant course objectives. The items' weak to moderate correlations with extraneous variables beyond the instructor's control (e.g., difficulty of subject matter, student work habits) provide evidence for divergent validity. Statistical details are discussed in the following sections.

Correlations Between Summary Measures and Progress on Relevant Objectives

Average student progress on relevant objectives (PRO) (i.e., objectives the instructor identifies as either "Important" or "Essential") has been and continues to be IDEA's best single measure of teaching effectiveness, because it reflects students' perceptions of how much they have learned. Students rate their progress on each of 12 learning objectives by responding 1 = *No apparent progress*, 2 = *Slight progress*, 3 = *Moderate progress*, 4 = *Substantial progress*, 5 = *Exceptional progress*. Using a sample of 17,183 classes from 105 institutions, Cashin and Downey (1992) found in separate regression analyses that the "excellent teacher" (r = .74) and "excellent course" (r = .77) items each accounted for more than 50% of the variance in PRO.

We replicated Cashin and Downey's (1992) study using student ratings collected in 681,715 classes from 2002 to 2013. Ratings of excellence of the instructor explained 64% (r = .80) of the variance in PRO; ratings of the course accounted for 67% (r = .82). The zero-order correlation between excellence of the instructor

[1] *Diagnostic Feedback* was previously called the "Diagnostic Form" or "Long Form"; *Learning Outcomes* was called the "Short Form."

and excellence of the course was .86. Employing hierarchical multiple regression analysis, we entered the extraneous variables (difficulty of subject matter, student motivation, student work habits, and class size) on the first step (R^2 = .247), followed by ratings of the instructor and course excellence on the second step (R^2 Change = .492) (see Table 1). The results corroborate Cashin and Downey's conclusion that, "because global items accounted for a substantial amount of the variance [in PRO], a short and economical form would capture much of the information needed for summative evaluation and longer diagnostic forms could be reserved for teaching improvement" (p. 563).

Table 1

Hierarchical Multiple Regression Analysis Predicting Average Student Progress on Relevant Objectives from Extraneous Variables and Average Ratings of the Instructor and Course

Predictor	R^2	β	*t*	*p*
Step 1	.247			
Difficulty of subject matter		.002	2.07	.038
Student motivation		.303	269.32	.001
Student work habits		.284	245.95	.001
Class size		-.081	76.98	.001
Step 2	.739			
Difficulty of subject matter		.084	129.62	.001
Student motivation		0	<1	.818
Student work habits		.154	222.04	.001
Class size		-.032	50.85	.001
Excellence of instructor		.407	312.61	.001
Excellence of course		.417	279.03	.001

Correlations Between Summary Measures and Extraneous Variables

Prior to the development of *Teaching Essentials*, extraneous student variables (i.e., motivation and work habits) and course variables (i.e., difficulty of subject matter and size of class) were used to adjust ratings on the two overall summary measures. Using a 5-point Likert scale (1 = *Definitely False* to 5 = *Definitely True*), students described the extent to which they *really wanted to take this course regardless of who taught it* (i.e., motivation) and their typical work habits (*As a rule, I put forth more effort than other students on academic work*). They rated the *difficulty of the subject matter*, using a 5-point scale (1 = *Much Less than Most Courses*, 2 = *Less than Most Courses*, 3 = *About Average*, 4 = *More than Most Courses*, and 5 = *Much More than Most Courses*). Instructors indicated the number of students enrolled (i.e., class size) on the *Faculty Information Form*.

Using the 2002-2013 research database, we computed Pearson *r* correlations between the extraneous variables and the two overall summary measures.[2] As shown in Table 2, subject matter difficulty is only trivially related to student perceptions of the instructor and the course (*r* = -.03 and -.01, respectively). However, student work habits, motivation, and class size are more strongly related to the two summary measures. Generally, ratings of the instructor and the course are somewhat higher in smaller classes and classes where students report higher than average work habits and motivation. Consequently, IDEA controls for the influence of these extraneous variables by computing adjusted scores on the two summary measures.

[2] The research database excludes classes with fewer than 10 student responses and courses taught by novice (first-year) users of IDEA. No single institution accounts for more than 5% of the database.

Table 2
Pearson r Correlations Between IDEA Summary Measures and Extraneous Variables

Variable	1	2	3	4	5	6	7	8
1. PRO	1.00							
2. Difficulty of subject matter	.10	1.00						
3. Student motivation	.41	.10	1.00					
4. Student work habits	.40	.25	.35	1.00				
5. Student background	.45	—	.48	.50	1.00			
6. Class size	-.12	.02	-.09	-.06	-.09	1.00		
7. Excellence of instructor	.80	-.03	.29	.21	.40	-.10	1.00	
8. Excellence of course	.82	-.01	.54	.32	.49	-.11	.86	1.00

Note. Correlation coefficients between student background and other variables, calculated using the 2002-2013 *Short Form* dataset (n = 226,838), are significant at $p < .01$. The correlation between student background and difficulty of subject matter is unavailable because the two items are not presented on the same instrument. All other coefficients are significant at $p < .001$.

Generalizability of the Adjusted Score Formulas

Class size and student ratings of work habits and motivation have long been used as extraneous variables in the IDEA adjusted score formulas. The student background item, *My background prepared me well for the course's requirements*, had appeared for a number of years on the *Short Form* but had not been used for adjusted scores. The past few decades of research in cognitive psychology and education has demonstrated that student background knowledge plays a critical role in student learning. Consequently, IDEA research staff set out to test its influence on student ratings.

Hoyt and Lee (2003) performed a preliminary analysis to determine whether the background variable should be included in adjusted score formulas. They found that it explained significant additional variance in student ratings of the two overall summary measures beyond that explained by work habits and motivation. Using ratings from 2006 to 2010, Benton and colleagues (Benton, Li, Brown, Guo, & Sullivan, 2015) tested whether work habits and motivation explained significant variance in the student background variable. If the amount of variance explained was large, then not much would be gained by adding student background as an extraneous variable. They found work habits and motivation explained 33% of the variance in student background preparation, which left 67% of the variance unexplained. Much could be gained, then, by adding student background to the adjusted score models.

Consistent with Hoyt and Lee's (2003) work, Benton et al. (2015) confirmed that student background added significant variance beyond work habits and motivation to adjusted score formulas. Taken together, the combination of extraneous variables (student work habits, motivation, and background) explained 24% of the variance in ratings of excellence of the instructor and 45% in overall ratings of the course. The bivariate correlations between student background and the two overall summary measures were moderate in strength (r = .40 and .49, respectively).

Benton et al. (2015) then applied cluster analysis and principal components analysis to test whether the regression models developed for adjusted scores differed across 38 discipline groups. No distinction could be made between discipline groups in the intercepts and slopes found to be significant in the regression models. Furthermore, no clear distinction could be made between disciplines in the adjusted scores produced from those models. On *Teaching Essentials*, then, average ratings on both the "excellent teacher" and "excellent course" items are adjusted for student work habits, motivation, background, as well as class size.[3]

[3] In the initial rollout of Teaching Essentials, adjustments were made only for student work habits, motivation, and class size until such time that adequate data could be collected to adjust for student background.

Relationships Between Summary Measures and Teaching Methods

On the previous version of the DF, students rated how frequently their instructor used each of 20 teaching methods (1 = *Hardly Ever*, 2 = *Occasionally*, 3 = *Sometimes*, 4 = *Frequently*, 5 = *Almost Always*). To investigate which methods were most highly correlated with each overall summary measure, Benton et al. (2015) applied Bayesian Model Averaging (BMA). BMA is an ensemble technique that tests multiple models to obtain better predictive performance than what could be obtained with a single model (Hoeting, Madigan, Raftery, & Volinsky, 1999). Separate analyses were conducted on overall ratings of the instructor and the course.

Seven of the 20 teaching methods were significantly related to either one or both of the overall summary measures. The results were generally consistent across small (10-14 students), medium (15-34 students), large (35-49 students), and very large (50 or more students) classes. Table 3 shows which teaching methods are associated with each overall measure, categorized by four major areas: organization, clarity, enthusiasm/expression, and rapport/interactions. Across multiple factor-analytic studies of student ratings (Braskamp & Ory, 1994; Feldman, 1989; Hativa, Barak, & Simhi, 2001; Marsh, 1987; Murray, 1997), those four broad teacher behaviors were highly correlated with ratings of teaching effectiveness (as summarized in Hativa, 2013). The seven teaching methods in Table 3 were considered most important, then, for making improvements in ratings of the instructor and the course, and they therefore all appear on *Teaching Essentials*.

Table 3
Teaching Methods Related to Overall Summary Measures

	Overall summary measure	
Teaching method category	Excellence of instructor	Excellence of course
Organization		6. Made it clear how each topic fit into the course
Clarity	10. Explained course material clearly and concisely	10. Explained course material clearly and concisely
Enthusiasm/expression	13. Introduced stimulating ideas about the subject	13. Introduced stimulating ideas about the subject 15. Inspired students to set and achieve goals which really challenged them 4. Demonstrated the importance and significance of the subject matter
Rapport/interactions	1. Displayed personal interest in students and their learning 2. Found ways to help students answer their own questions	

Comparing Ratings Collected Online Versus on Paper

Responses to IDEA SRI are very similar regardless of whether they are administered online or on paper (Benton, Webster, Gross, & Pallett, 2010b). Table 4 presents means and standard deviations, along with Cohen's *d*, for the two summary measures collected from the DF either on paper (*N* = 651,587) or online (*N* = 53,000). Data came from ratings administered during the years 2002 to 2008. Cohen (1988) considered effect sizes approximating .20 as small, .50 as medium, and .80 as large. The differences in Table 4 are trivial, indicating average ratings of the instructor and the course do not differ meaningfully between paper and online formats.

Table 4
Means and Standard Deviations for Summary Measures Administered on Paper and Online, 2002-2008

	Paper		Online		
Summary Measure	*M*	*SD*	*M*	*SD*	Cohen's *d*
Excellence of instructor	4.25	0.61	4.16	0.62	.15
Excellence of course	4.01	0.59	3.98	0.59	.05

The correlations between the seven teaching methods and the two overall summary measures are also very similar whether administered on paper or online. The Pearson *r* correlations presented in Table 5 provide strong evidence for the generalizability of relationships between teaching methods and summary measures across survey delivery formats.

Comparing Ratings in Face-to-Face and Online Courses

Student ratings also tend to be highly similar between face-to-face and online courses (Benton, Webster, Gross, & Pallett, 2010a). Using data collected from 2002 to 2008, Benton et al. (2010a) compared instructors who had administered the DF online and had taught the course exclusively either face-to-face (*N* = 5,272) or online (*N* = 13,416). Table 6 presents descriptive statistics for scores on the two overall summary measures by type of course. Cohen's *d* values indicated there were no meaningful differences between course formats.

Table 7 (*see page 6*) presents correlations between the seven teaching methods and the two overall summary measures by type of course. As was the case with paper versus online administration, no meaningful differences were found in the direction or magnitude of the correlations.

Table 5
Correlations Between Teaching Methods and Overall Summary Measures for Ratings Administered on Paper versus Online, 2002-2008

	Excellence of instructor		Excellence of course	
Teaching method	Paper	Online	Paper	Online
Displayed interest	.85	.87	.74	.74
Helped students answer their own questions	.86	.89	.77	.78
Demonstrated the importance of subject	.83	.86	.80	.81
Made it clear how each topic fit into the course	.85	.86	.80	.81
Explained material clearly	.90	.91	.80	.83
Introduced stimulating ideas	.83	.86	.82	.83
Inspired students	.75	.81	.75	.77

Table 6
Means and Standard Deviations for Overall Summary Measures in Face-to-Face and Online Courses

	Face-to-face		Online		
Overall rating	*M*	*SD*	*M*	*SD*	Cohen's *d*
Excellence of instructor	4.20	0.72	4.18	0.71	.03
Excellence of course	4.05	0.69	4.06	0.68	.01

Table 7
Correlations Between Teaching Methods and Overall Summary Measures for Ratings in Face-to-Face versus Online Courses

	Excellence of instructor		Excellence of course	
Teaching method	F2F	Online	F2F	Online
Displayed interest	.83	.85	.69	.72
Helped students answer their own questions	.85	.83	.73	.71
Demonstrated the importance of subject	.83	.82	.76	.75
Made it clear how each topic fit into the course	.82	.79	.75	.74
Explained material clearly	.87	.85	.76	.77
Introduced stimulating ideas	.84	.79	.76	.74
Inspired students	.78	.77	.71	.71

Note. F2F = face-to-face.

Summary of Validity Evidence

Overall summary ratings of the instructor and the course are more strongly related to student progress on relevant objectives than they are to the extraneous variables of student work habits, motivation, background, and class size. This means that student impressions of the instructor and the course have more to do with self-reported progress on important learning outcomes than with factors beyond the instructor's control. Even so, *Teaching Essentials* adjusts scores to control for the effects of extraneous variables on overall summary ratings. The models used to adjust ratings are not distinguishable across discipline groups.

The seven teaching methods adopted for *Teaching Essentials* align with four broad categories of teacher behaviors highly correlated with ratings of teaching effectiveness. The relationships between teaching methods and the two summary measures are consistent across classes that vary in size, in how the survey is administered (i.e., paper vs. online), and in the educational setting (i.e., face to face vs. online). Moreover, average ratings on the two summary measures do not differ meaningfully by survey administration or course setting.

RELIABILITY

Reliability evidence is important for determining whether student ratings are consistent enough to be used as a source of evidence for making judgments about teaching effectiveness. When ratings vary substantially among students within the same course or when average instructor ratings change dramatically from one class to another, evaluative decisions about effectiveness are difficult. As revealed in the following paragraphs, credible evidence exists to support both the class-level and instructor-level reliability of each item on *Teaching Essentials*.

Class-level Reliability

Hoyt and Lee (2002) provided evidence for adequate class-level reliability (i.e., consistency in ratings by students in the same class) by computing split-half reliabilities on each of the items on the DF. Classes with the number of student respondents ranging from 13-17 were randomly split and means were computed for each half. The means were then correlated, and the Spearman-Brown formula was applied to estimate reliabilities for class size ranges of 10-14, 15-34, 35-49, and 50+. For all items, split-half reliability estimates were above .80 when class size was at least 15. Standard errors of measurement (SEM) were approximately .30 or less once class size reached 10.

Instructor-level Reliability

Consistency in ratings within the same class is a prerequisite of instructor-level reliability (i.e., stability in ratings of the same instructor across different classes) (Gillmore, 2000). However, ratings can have adequate class-level reliability without being consistent at the instructor-level. Benton et al. (2015) obtained measures of instructor-level reliability on IDEA items by computing inter-class reliability coefficients on a subset of data from 2,500 instructors who had been rated in at least five classes. The Spearman-Brown prophecy formula was then applied to estimate reliabilities for 1 to 15 classes. All reliability estimates for *Teaching*

Essentials items approached or exceeded .60 for a single class. When at least two classes were rated, all coefficients were above .70 and most approached or exceeded .80. Reliability coefficients increased as the number of classes rated increased; all were .90 or greater when at least seven classes were rated. Standard errors averaged .30 or less for all *Teaching Essentials* items when at least two classes were rated.

Summary of Reliability Evidence

Substantial evidence exists to support both class-level and instructor-level reliability for rating items on *Teaching Essentials*. As class size and the number of classes rated increases, faculty and administrators can gain greater confidence in the reliability of the ratings.

INTENDED USE OF *TEACHING ESSENTIALS*

The primary intent of *Teaching Essentials* is to gather feedback from students to inform instructors about suggestions for making improvements in teaching and the course. For each course section rated, instructors receive a report that provides suggestions for actions they might take with respect to each teaching method. Two factors are considered in creating the decision rules for "recommended actions": (a) the magnitude of the difference between the instructor's average score on the teaching method and the mean rating of other courses that are similar in class size and level of student motivation, and (b) the percentage of students reporting the teaching method used by the instructor "frequently" or "almost always." IDEA uses this information to recommend whether each teaching method is a "strength to retain," whether the instructor should "retain current use or consider increasing," or whether the instructor should "consider increasing use."

Feedback from the *Teaching Essentials* report can be beneficial to both full- and part-time faculty. The teaching methods are general enough to be applicable to a variety of instructional approaches, class sizes, course settings, and disciplines. In most educational settings—whether instruction occurs online or face to face—being organized, explaining content clearly, introducing stimulating ideas, inspiring students to set and achieve goals, demonstrating the importance of the subject matter, displaying an interest in students, and helping students find ways to answer their own questions are helpful methods.

Teaching Essentials enables both end-of-course and periodic feedback. One can take a traditional approach of waiting until instruction is nearing its end to administer the survey. Recommended actions can then be taken to make formative decisions about improvement the next time the instructor teaches the course. In addition, scores from multiple classes on the two overall measures, adjusted for extraneous variables, can be used *in combination with other information* to make summative decisions about the instructor and the course. However, *Teaching Essentials* collects no information about average student progress on relevant course objectives, which IDEA considers the best single measure of teaching effectiveness.

Alternatively, because *Teaching Essentials* contains only 12 items, the instructor could conduct periodic assessments multiple times during the course. Student intermittent feedback could signal which adjustments to make while the course is in progress. Effects of the adjustments could then be measured if the instrument is administered again later in the semester. In this way the current students could benefit from instructional improvements.

A clear advantage of *Teaching Essentials* is suitability with mobile technology. Students can complete the ratings conveniently in class on smart phones or tablets, which could potentially increase response rates.

References

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

Benton, S. L., & 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., Li, D., Brown, R., Guo, M., & Sullivan, P. (2015). *IDEA Technical Report No. 18: Revising the IDEA Student Ratings of Instruction System, 2002-2011 data*. Manhattan, KS: The IDEA Center.

Benton, S. L., Webster, R., Gross, A. B., Pallett, W. (2010a). *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.

Benton, S. L., Webster, R., Gross, A. B., Pallett, W. (2010b). *IDEA Technical Report No. 16: An analysis of IDEA Student Ratings of Instruction using paper versus online survey methods, 2002-2008 data*. Manhattan, KS: The IDEA Center.

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

Cashin, W. E., & Downey, R. G. (1992). Using global student rating items for summative evaluation. *Journal of Educational Psychology*, *84*, 563-572.

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education*, *30*, 583-645.

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. (2013). *Student ratings of instruction: Recognizing effective teaching*. Oron Publications.

Hativa, N., Barak, R., & Simhi, E. (2001). Exemplary university teachers: Knowledge and beliefs regarding effective teaching dimensions and strategies. *Journal of Higher Education, 72*, 699-729.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial with discussion. *Statistical Science, 14*, 382-417.

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., & Lee, E. (2003, April). *Short form extraneous variable analysis*. Manhattan, KS: The IDEA Center.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253-388.

Murray, H. G. (1997). Effective teaching behaviors in the college classroom. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 171-204). New York, NY: Agathon Press.

T: 800.255.2757
T: 785.320.2400

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