IDEA PAPER #50

Insight.Improvement.Impact.®

Student Ratings of Teaching: A Summary of Research and Literature

Stephen L. Benton,[1] The IDEA Center
William E. Cashin, Emeritus professor • Kansas State University

"Ratings of overall effectiveness are moderately correlated with independent measures of student learning and achievement. Students of highly rated teachers achieve higher final exam scores, can better apply course material, and are more inclined to pursue the subject subsequently." (Davis, 2009, p. 534)

This IDEA Paper is an update of IDEA Paper No. 32 *Student Ratings of Teaching: The Research Revisited* (Cashin, 1995). Much of the content of IDEA Paper No. 32 is retained where no subsequently published study has changed its basic conclusions. However, studies or reviews of the literature that provided questions, modifications, or further support for its conclusions were included in this paper. We have attempted to summarize the conclusions of the major reviews of the student ratings research and literature from the 1970s to 2010. That literature is extensive and complex; a paper this brief can offer only broad, general summaries and limited citations.

At the end of 2010, there were 2,875 references in the ERIC database using the descriptor "student evaluation of teacher performance," the ERIC descriptor for student ratings of teaching /student evaluations of teaching (SRT/SET). By adding the descriptor "higher education," the number was reduced to 1,852. Restricting our search to the years 1994 to 2010 yielded 542 references. No major summary of the student ratings research was found in those 542 references, only specific studies. However, ERIC no longer included chapters from the annual *Higher Education: Handbook of Theory and Research* or compilations of chapters from *Effective Teaching in Higher Education: Research and Practice* (Perry & Smart, 1997) or *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective* (Perry & Smart, 2007).

We found especially useful the following chapters published in the book by Perry and Smart (2007); Abrami, Rosenfeld, and Dedic, (2007); Abrami, d'Apollonia, Rosenfeld (2007); Feldman (2007); Marsh (2007); Murray (2007); Theall and Feldman (2007). Those interested are encouraged to read these reviews and their individual references for more details. For readers with less time, Davis (2009), Forsyth (2003), Svinicki and McKeachie (2011), and Wachtel (1998), as well as earlier works by Braskamp and Ory (1994) and Centra (1993), have sections summarizing student-ratings research.

Although the ERIC descriptor for student ratings is "student evaluation of teacher performance," we prefer the term "student ratings." Whereas "evaluation" has a definitive and terminal connotation of determining worth, "ratings" refer to data that need interpretation. Using the term "rating" rather than "evaluation" helps to distinguish between the people who provide the information (sources of data) and those who interpret it (evaluators) in combination with other sources of information. Viewing student ratings as data rather than as evaluations puts them in their proper perspective.

Writers on faculty evaluation are almost universal in recommending the use of multiple sources of data. No single source of information – including student ratings – provides sufficient information to make a valid judgment

[1]*Authors listed in alphabetical order.*

about an instructor's *overall teaching effectiveness*. Further, there are important aspects of teaching that students are not competent to rate. For elaborations on this issue, see IDEA Paper No. 21, *Defining and Evaluating College Teaching* (Cashin, 1989; see also Abrami, d'Apollonia, & Rosenfeld, 2007; Abrami, Rosenfeld, & Dedic, 2007; Arreola, 2006; Braskamp & Ory, 1994; Cashin, 2003; Centra, 1993; Davis, 2009; Forsyth, 2003; Marsh, 2007; Svinicki & McKeachie, 2011).

Persistent Misconceptions About Student Ratings

Several authors have pointed out misconceptions about student ratings that are *un*supported by research and that make improved practice difficult (Aleamoni, 1987; Feldman, 2007; Kulik 2001; Svinicki & McKeachie, 2011; Theall & Franklin, 2007). The following are some of the most commonly held misconceptions:

- Students cannot make consistent judgments.
- Student ratings are just popularity contests.
- Student ratings are unreliable and invalid.
- The time of day the course is offered affects ratings.
- Students will not appreciate good teaching until they are out of college a few years.
- Students just want easy courses.
- Student feedback cannot be used to help improve instruction.
- Emphasis on student ratings has led to grade inflation.

These myths ignore more than 50 years of credible research on the validity and reliability of student ratings. They persist, unfortunately, largely due to ignorance of the research, personal biases, suspicion, fear, and general hostility toward any evaluation process (Theall & Feldman, 2007).

In the sections that follow, we briefly summarize research on several aspects of student ratings that provide evidence of their reliability and validity. We also examine extraneous student, instructor, and course characteristics that are either unrelated or related to student ratings. Finally, we address ratings administered online versus on paper, ratings in online versus face-to-face courses, and the usefulness of student ratings.

Reliability

Reliability refers to the consistency, stability, and generalizability of measurement data. With respect to student ratings, reliability most often concerns *consistency* or interrater agreement (that is, within a given class whether students tend to give similar ratings on a given item). Reliability coefficients typically range from .00 to 1.00 with higher values indicating greater consistency. Standard errors of measurement (SEM), which are sometimes reported, indicate the amount of error or spread (+ or -) in the scores. Reliability estimates vary depending upon the number of raters. Generally, the more raters, the more reliable or dependable are scores based on those ratings. As an example, in the IDEA system (Hoyt & Lee, 2002a) the average split-half reliabilities and SEMs for the student-rating items are broken down by class sizes in Figure 1. The coefficients in Figure 1 show that as class size increases, reliability (or consistency in the scores) increases, but the amount of error (SEM) decreases. So, error is the flipside of reliability.

Figure 1 • Average Split-half Reliabilities and Standard Errors of Measurement by Class Size for IDEA Student Ratings.

Class Size	Reliability	SEM
10-14 students	.78	.27
15-34 students	.87	.21
35-49 students	.92	.16
50+ students	.94	.14

Note. SEM = standard error of measurement

Similar estimates are typically found with other well-designed forms (i.e., forms developed with the assistance of someone knowledgeable about educational measurement and the research on student ratings of teaching). As a general rule, multiple classes provide more reliable results than a single class. When ratings are based on fewer than 10 students, multiple class ratings are especially important.

Stability is concerned with agreement between raters *over time*. In general, ratings of the same instructor across semesters tend to be similar (Braskamp & Ory, 1994; Centra, 1993). In a longitudinal study, Overall and Marsh (1980) compared end-of-course ratings with ratings by the same students a year or more later (at least one year after graduation). The average correlation was .83.

Generalizability refers to how accurately the data reflect the instructor's *general* teaching effectiveness, not just how effective he or she was in teaching a particular course in a given term. Marsh (1984) addressed this question by categorizing student ratings data from 1,364 classes into four categories: the same instructor teaching the same course but in different semesters, the same instructor teaching a different course, different instructors teaching the same course, and different instructors teaching different courses. This permitted him to study the differential effects of the instructor and the course. He then correlated student ratings in the four different categories, separating items related to the instructor (e.g., student ratings of the instructor's enthusiasm, organization, discussion) from student background items (e.g., student's prior subject interest, reasons for taking the course). The average correlations are shown in Figure 2. The instructor-related correlations were higher for the same instructor, even when teaching a different course. The correlations for the background items – more tied to

the course than the instructor – were higher for the same course. Marsh concluded, therefore, that the instructor, *not* the course, is the *primary* determinant of students' ratings. Marsh's results are comparable to those found in other generalizability studies (Gillmore, Kane, & Naccarato, 1978; and Hogan, 1973).

Figure 2 • Average Correlations among Different Sets of Classes for Student Ratings of Instructor and Background Characteristics.

	Course			
	Same		Different	
Instructor	Instructor Items	Background Items	Instructor Items	Background Items
Same	.71	.69	.52	.34
Different	.14	.49	.06	.21

Generalizability is especially relevant when making personnel decisions about an instructor's general teaching effectiveness. Keeping in mind such decisions should be based on additional information beyond student ratings (see Cashin, 2003), we offer the following guidelines. If the instructor teaches only one course (e.g., part-time instructors), then consistent ratings from two different terms may be sufficient. For most instructors, however, ratings from a variety of courses are necessary, preferably two or more courses from every term, for at least two years, totaling six to eight courses. If there are fewer than 10 raters in any of the classes, data from additional classes are recommended.

Validity

In educational measurement, the basic question related to validity is: Does the test – the variable – measure what it is supposed to measure? For student ratings this translates into: To what extent do student rating items measure some aspect of teaching effectiveness? Unfortunately there is no agreed upon definition of "effective teaching" or any single, all-embracing criterion (see, for example, Cashin, 2003). The best one can do is to try various approaches, collecting data that either support or contest the conclusion that student ratings reflect effective teaching.

As is the case with reliability, validity is *not* a characteristic inherent in a student ratings instrument. Validity is determined by how the ratings are used – how they are interpreted and what actions follow from those interpretations – referred to as the consequential basis of validity (Messick, 1989). McKeachie (1997) cautioned that faculty and administrators need education about how to use ratings appropriately (i.e., validly).

Student ratings typically serve several purposes. They help faculty improve their teaching and courses, administrators make decisions about salary and promotion, committee members select teaching award winners, institutions conduct program reviews, and students select courses. When used in combination with other measures of teaching effectiveness, ratings can serve all of these purposes. However, when used for unintended purposes (e.g., basing course content on student-rating form content, making administrative decisions based on ultra-fine discriminations in ratings, and altering standard administration procedures), validity is threatened (Ory & Ryan, 2001).

Researchers have traditionally taken one of several approaches to validity studies: (a) correlating ratings in multiple sections of the same course with student achievement on a common examination; (b) correlating ratings with other criteria (e.g., alumni-, peer-, or self-ratings); (c) examining bias by correlating ratings with student, instructor, and course characteristics; (d) manipulating administrative procedures; (e) conducting experiments in non-natural settings; and (f) analyzing the underlying dimensions of ratings (Ory & Ryan, 2001). Evidence from all such studies affects the meaning and interpretation of student ratings or their *construct* validity (Messick, 1995). In the paragraphs that follow, we summarize research employing each of these approaches.

Validity Approach One: Correlating Student Ratings with Achievement

Theoretically, the best indicant of effective teaching is student learning. Other things being equal, the students of more effective teachers should learn more. A number of studies have attempted to examine this hypothesis by comparing multiple-section courses. For example, Benton and colleagues (Benton, Duchon, & Pallett, 2011) examined student ratings in multiple sections of the same course taught by the same instructor. They correlated student ratings of progress on objectives the instructor identified as relevant to the course (using IDEA student ratings) with their performance on exams tied to those objectives. Student ratings correlated positively with four out of five exams and with the course total points ($r = .32$). In contrast, student ratings of progress on objectives the instructor considered of minor or no importance were not related to exam performance.

Other studies have been conducted on multiple instructors who teach different sections of the same course. The instructors use the same syllabus and textbook and, most importantly, the same *external* final exam (i.e., an exam developed by someone *other* than the instructors). Student ratings of the course and instructor are then correlated with final exam scores. Cohen (1981, 1987) and Feldman (1989b) reviewed several studies of this kind and, for each one, correlated final exam scores with various student ratings items.[2] Figure 3 presents the

[2]*The authors converted various summary statistics reported in the multi-section studies into Pearson-product moment correlations.*

average correlations as they were reported in Cohen (1981, 1987) and Feldman (1989b). Both Cohen's and Feldman's correlational approaches were consistent in identifying the instructional dimensions (e.g., teacher preparation and course organization, teacher clarity, teacher stimulation of student interest, and students' perceived impact or outcome of the course) most highly correlated with student achievement. (See also Abrami, 2001, and Kulik, 2001, for support of the relationship between student learning and student ratings.)

Figure 3 • Correlations between Student Final Exam Performance and Various Dimensions of Student Ratings.

Student Ratings of:	Average Correlations with Final Exam Across Three Studies		
	Cohen (1981)	Cohen (1987)	Feldman (1989b)
Achievement/learning	.47	.39	.46
Overall course	.47	.49	–
Overall instructor	.44	.45	–
Teacher skill:	.50	.50	–
-course preparation	–	–	.57
-clarity of objectives	–	–	.35
Teacher structure:	.47	.55	–
-understandableness	–	–	.56
Teacher rapport:	.31	.32	–
-availability	–	–	.36
-respect for students	–	–	.23
Teacher interaction:	.22	.52	–
-encouraging discussion	–	–	.36
Evaluation	–	.30	–
Feedback	–	.28	–
Interest/motivation	–	.15	–
Difficulty	–	-.04	–

In a follow-up study, Feldman (2007, pp. 104-105) reported the average correlations between a measure of student achievement and 24 specific instructional dimensions often measured by specific student rating items. In a separate table, Feldman (2007, pp. 112-113) also compared the correlations of various instructional dimensions with *student achievement* and students' *overall evaluation of the teacher*. The correlations with achievement and overall evaluations of teaching were not always of the same magnitude (e.g., quality and frequency of feedback correlated only .23 with student achievement but .87 with overall evaluation), but they showed the positive contribution of various instructional dimensions to both outcomes.

With respect to IDEA student ratings of instruction, several dimensions of teaching are strongly related ($r > .80$) to overall global ratings of the teacher: explaining course material clearly and concisely, finding ways for students to answer their own questions, displaying personal interest in students and their learning, making it clear how each topic fits into the course, demonstrating the importance of the subject matter, and introducing stimulating ideas about the subject (Hoyt & Lee, 2002a). Furthermore, the five IDEA teaching approaches (*Stimulating Student Interest, Fostering Student Collaboration, Establishing Rapport, Encouraging Student Involvement, and Structuring the Classroom*) together explain 85 percent of the variance in the "excellent teacher" item. Instructors wanting to increase their global teacher ratings should focus improvement efforts especially on enhancing their communication, motivational, and rapport-building skills (IDEA Research Note 1, 2003).

The correlations reported in this section are far from perfect, in part because some of the variables that correlate with student learning are related to student characteristics (e.g., ability or motivation). In addition, college exams typically have less than perfect reliability, which attenuates the correlations. Nonetheless, the multi-section studies show that classes in which the students gave the instructor higher ratings tended to be the ones where the students learned more (i.e., scored higher on the external exam).

Validity Approach Two: Correlating Ratings with Other Criteria

Instructor self-ratings. Researchers have sought a criterion of effective teaching acceptable to faculty. One possibility is self-ratings completed by the instructor, often using the same instrument used by students. In a review of 19 studies, Feldman (1989a) reported an average correlation of .29 between instructor self-ratings and student ratings. In another study by Marsh and colleagues (Marsh, Overall, & Kesler, 1979), instructors were asked to rate their teaching effectiveness in two courses in order to see if the course the instructor rated highest was also rated highest by the students. Student ratings were indeed higher in the courses the instructors indicated were more effectively taught. The median correlation – across six factor scores – was .49 between the instructor and student ratings. In a related study, Marsh (1982) found that 34 of the 35 correlations between student ratings and instructor self-ratings were statistically significant, with a median correlation of .30. Subsequently, Marsh and Dunkin (1997) found a median correlation of .45 between instructor self-ratings and student ratings on nine scale scores. Such findings support the criterion-related validity of student ratings.

In spite of the consistent positive correlations between student ratings and instructor self-ratings, some might still question whether students have an appropriate view of effective teaching. To address this concern, Feldman

(1988) reviewed 31 studies and found that students' views of effective teaching were very similar to the instructor's view (the average correlation was .71). However, some subtle differences in emphasis existed. Students tended to assign more importance to the instructor being interesting, having good speaking skills, and being available to help; students also focused more on the outcomes of instruction (e.g., what they learned). In contrast, instructors placed relatively more emphasis on challenging and motivating students, setting high standards, and fostering student self-initiated learning.

Feldman's (1988) findings do not necessarily indicate students undervalue instructors who set high standards and foster student self-initiated learning. Using IDEA student ratings, Hornbeak (2009) found that students' desire to take a course from an instructor was positively correlated (r = .54) with how much the instructor expected students to take their share of responsibility for learning. Moreover, students have a stronger desire to take a course (r = .52) when they rate the instructor as having high achievement standards (Hoyt & Lee, 2002a).

Ratings by administrators. Student ratings correlate moderately with administrator ratings of the instructor's general reputation, as coefficients range from .47 to .62 (Kulik & McKeachie, 1975). Feldman (1989a), using global items, found a lower average correlation of .39.

Ratings by colleagues. Instructor ratings by colleagues that are *not* based on classroom observations are moderately correlated with student ratings, r of .48 to .69 (Kulik & McKeachie, 1975). Feldman (1989a) found an average correlation of .55, using global ratings. However, ratings by colleagues can be unreliable and uncorrelated with student ratings when made by untrained observers in single classroom visitations employing an unsystematic approach (i.e., different faculty visiting the same class tend to disagree) (Marsh, 2007; Marsh & Dunkin, 1997).

Ratings by alumni. Some faculty may question whether current students can adequately judge the long-term effects of instruction. However, end-of course student ratings are positively correlated with retrospective ratings of an instructor provided by the same students several years later, r of .54 to .80 (Braskamp & Ory, 1994). Overall and Marsh (1980) and Feldman (1989a) reported average correlations of .83 and .69, respectively. These findings belie the conventional wisdom that students only come to appreciate teaching *after* they graduate and enter into the real world as working adults.

Ratings by trained observers. A few studies have used external observers who were trained to make classroom observations (see Feldman, 1989a; also Marsh & Dunkin, 1992). Reviewing five studies, Feldman reported an average correlation of .50 between the ratings of trained observers and global student ratings. In a related study, Murray (1983) reported a median reliability of .76 among the ratings of trained observers, which suggests ratings by colleagues might be more reliable if faculty were trained prior to making classroom observations.

Student comments. Some faculty members may question the value of student objective rating scales, giving preference to student written comments to open-ended questions. In one study of 14 classes, Ory and colleagues (Ory, Braskamp, & Pieper, 1980) found a correlation of .93 between a global instructor item and students' written comments. In a second study of 60 classes, the authors (Braskamp, Ory, & Pieper, 1981) found a correlation of .75. More recently, Burdsal and Harrison (2008) found a correlation of .79 in a sample of 208 classes. These studies suggest that, *for personnel decisions*, the information from student ratings considerably overlaps the information in student comments. Nonetheless, when decisions are made about promotion, faculty generally regard written comments as less credible than student responses to objective comments. On the other hand, faculty rate written comments as more credible when the purpose is for self-improvement (Braskamp et al., 1981).

The studies cited thus far provide evidence of the validity of student ratings. Student ratings are significantly and consistently related to student achievement, teacher self-ratings, administrator and colleague ratings, ratings by trained observers, and student written comments. In the next section, we consider possible biases in student ratings.

Validity Approach Three: Examining Possible Sources of Bias

One need not talk with faculty very long to be aware of their concern about possible biases in student ratings. Some writers have suggested that bias can be defined as anything *not under the control of the instructor*. However, Marsh (2007) offered another definition: "Bias exists when a student, teacher, or course characteristic affects the evaluations made, either positively or negatively, *but is* un*related to any criteria of good teaching*" (p. 350; see also Centra, 2003, p. 498). By this definition, the correlations between student ratings and class size, or between student ratings and student interest in the course, are not biases because students in small classes and students who are interested in the subject matter *actually do tend to learn more* and, hence, give their teachers higher ratings. Rather than using the term "bias," we distinguish between variables (when correlated with student ratings) that possibly require control versus those that do not require control, especially when making personnel decisions.

Variables *Not* Requiring Control. Despite widespread faculty concern, researchers have discovered relatively few variables that correlate with student ratings *but that are not* related to instructional effectiveness (i.e., student learning). Feldman (2007, pp. 97-98) listed five variables where *slightly* higher ratings were *sometimes* found. These included higher ratings for smaller versus larger classes,

lower- versus upper-level courses, higher- versus lower-ranked faculty, students taking elective versus required courses, and students in major versus non-major courses. He then discussed some factors that might account for these variables *not* being considered biases. For example, at certain institutions, higher-ranked faculty may, on average, be better teachers and thus deserve higher ratings. Teachers may also be less effective in large than small classes and, accordingly, receive lower ratings (i.e., not because students take out their disdain for large classes by assigning lower ratings). Generally, the following variables tend to show *little or no* relationship to student ratings and in our judgment do *not* require control:

A. Instructor variables not related to student ratings:

1. **Age and teaching experience.** In general, instructor age and years of teaching experience are not correlated with student ratings. However, where weak correlations have been found they tend to be negative (i.e., older faculty receive *lower* ratings, Feldman, 1983; Renaud & Murray, 1996). Marsh and Hocevar (1991) pointed out that most of the studies of these variables have been cross-sectional comparisons of faculty cohorts that represent different age groups. In a *longitudinal* study, Marsh and Hocevar (1991) analyzed student ratings of the *same* instructors across 13 years and found *no* systematic changes within instructors over time.

   Centra (2009) found that first-year teachers tend to receive lower ratings compared to experienced assistant professors and higher-ranked faculty. He concluded the lower ratings do not point to bias but probably reflect differences in teaching skills, because first-year faculty are most likely still learning how to teach.

2. **Gender of the instructor.** In a review of 14 laboratory or experimental studies (where students rated descriptions of *fictitious* teachers who varied in gender), Feldman (1992) found few gender differences in global ratings. However, in a few studies male teachers received higher ratings. In a second review of 28 studies of global ratings – *involving actual student ratings of real teachers* – Feldman (1993) found a very slight average correlation between instructor gender and student ratings (r = .02) that favored female instructors. Women also received slightly higher ratings on sensitivity and on concern with student level of preparedness and progress (r = .12). However, ratings of male and female teachers did not differ meaningfully on other dimensions of teaching.

   Some researchers have reported a student-gender by instructor-gender interaction. Feldman (1993) found, for example, that female students tended to give higher ratings to female teachers, and male students tended to assign higher marks to male instructors. Centra (2009) found that female instructors received slightly higher ratings, especially by female students, but that these were *not* accompanied by higher student self-ratings of learning. He, nonetheless, concluded that gender was not a bias, because the higher ratings might have reflected differences in teaching style: women were more likely to use discussion than lecture, and women appeared to be more nurturing to their students (i.e., possibly more student-centered). Regardless, the effect due to gender, although statistically significant, was so small that it would most likely not affect personnel decisions (see also Centra & Gaubatz, 2000).

3. **Race.** Centra (1993) found, as we did, few studies of student ratings and instructor race conducted in North America. He speculated that students of the same race as the instructor *might* rate the instructor higher. However, in a doctoral dissertation using IDEA, Li (1993) found *no* differences between Asian and American students in their global ratings of (presumably Caucasian) instructors.

4. **Personal characteristics.** Few personality traits have been found to correlate with student ratings (Braskamp & Ory, 1994; Centra, 1993). Using instructor self-report (e.g., personality inventories, self-description questionnaires) as a criterion measure, Feldman (1986) found only two (out of fourteen) traits that had average correlations with a global teaching item that approached practical significance: positive self-esteem (r = .30), and energy and enthusiasm (r = .27).

   Murray, Rushton, and Paunonen (1990) found significantly different patterns of correlations between personality traits and student ratings among psychology instructors teaching six different types of courses (e.g., introductory, graduate). They concluded that instructor personalities tended to be differentially suited to different types of courses. In a follow-up study using the same measures of personality, Renaud and Murray (1996) found positive correlations between average scores on a 10-item student ratings scale and colleagues' ratings of the instructor's orderliness (.65), defined as being neat and organized and disliking clutter and confusion. Working for the approval and recognition of others was also positively correlated (.56) with teaching effectiveness.

   What matters more than personality, however, is how the instructor's personal characteristics are manifested in the classroom. Most of the relationship between instructor personality and student ratings can be explained by the behaviors the instructor exhibits when teaching (Erdle, Murray, & Rushton, 1985). Put simply, the effect of instructors' personalities on ratings "may be caused more by what they do in their teaching than by who they are" (Braskamp & Ory, 1994, p. 180). We suggest that the personality traits associated with student ratings enhance the instructor's teaching effectiveness and should, therefore, not be controlled.

5. **Research productivity** has little correlation with student ratings (Centra, 1993). In his review, Feldman (1987) found an average correlation of .12 between research productivity and ratings of overall teaching effectiveness. This very low correlation suggests that research productivity is indicative of *neither* good *nor* bad teaching (i.e., due to more time being devoted to research). Marsh and Hattie (2002) reported similar results.

B. Student variables not related to student ratings:

1. **Age of the student** has little effect on student ratings (McKeachie, 1979; Centra 1993).

2. **Gender of the student.** Feldman (1977, 1993, 2007) reported no consistent gender effect, although some have reported a student-gender by instructor-gender interaction (see earlier section on instructor variables). In a comprehensive study of gender, Centra and Gaubatz (2000) analyzed actual student ratings (rather than data from simulations) across a large number of two- and four-year institutions, involving a variety of academic disciplines. They found some gender preferences, particularly female students for female instructors. Although the differences were statistically significant, they were not large and would most likely not impact personnel decisions. Centra and Gaubatz (2000) speculated that the higher ratings female instructors received from female students, and sometimes from male students, might have reflected preferences for certain teaching styles. Women in their study were more likely than men to use discussion than lecture, and they were more nurturing to students, as reflected in their scores on certain rating scales.

3. **Level of the student** (e.g., first year, senior) has little practical effect on ratings (McKeachie, 1979).

4. **Student GPA.** In a summary of research, Davis (2009) concluded there is little or no relationship between student ratings, GPA, and year in college, citing several authors (Abrami, 2001; Brashkamp & Ory, 1994; Centra, 1993; Marsh & Dunkin, 1992 [see also Marsh & Dunkin, 1997]; Marsh & Roche, 2000; and McKeachie, 1997).

5. **Student personality.** No meaningful relationships exist between student personality and ratings (Abrami, Perry, & Leventhal, 1982).

***C.* Course variables *not* related to student ratings:**

1. **Time of day** the course is taught has no meaningful influence on student ratings (Aleamoni, 1981; Feldman, 1978).

D. Administrative variables not related to student ratings:

1. **Time during the term when ratings are collected.** Any time during the second half of the term seems to yield similar results (Feldman, 1979). Costin (1968) found no difference in ratings administered at the end versus the middle of the semester. Carrier et al. (1974) found no difference between ratings administered the last week versus the day of the final examination (although IDEA recommends against administering ratings at that time). Finally, Frey (1976) found no difference in ratings administered the last week of class versus the first week of the next semester.

Variables *Possibly* Requiring Control. The research cited thus far suggests that many variables suspected of biasing student ratings are *not* correlated with them to any practically significant degree. However, research suggests the following variables are correlated with student ratings and may require control.

A. Instructor variables related to student ratings:

1. **Faculty rank.** Regular faculty members tend to receive higher ratings than graduate teaching assistants (Braskamp & Ory, 1994). This variable may *NOT* require control because regular faculty as a group are more experienced and, therefore, tend to be more effective teachers than do graduate teaching assistants.

2. **Expressiveness.** The Dr. Fox effect – where a professional actor, who delivered a dramatic lecture but with little meaningful content, received high ratings – suggested that student ratings might be influenced more by an instructor's style of presentation than by the substance of the content (Naftulin, Ware, & Donnelly, 1973). The literature generated by the Dr. Fox study was complex (see Abrami, Leventhal, & Perry, 1982) but was clarified in the findings of Marsh and Ware (1982). They found that when student extrinsic motivation to achieve in a course is low, the influence of instructor expressiveness is substantial. Being more expressive produces higher student ratings *and* higher examination performance. More specifically, manipulations of instructor expressiveness primarily influence ratings of instructor enthusiasm; manipulations of lecture content primarily influence ratings of instructor knowledge, as well as student exam performance. In short, making the class interesting as well as informative helps students pay attention. Expressiveness, therefore, tends to enhance learning and we suggest does *NOT* require control.

B. Student variables related to student ratings:

1. **Student motivation.** Instructors are more likely to receive higher ratings in classes where students had a prior interest in the subject matter (Marsh & Dunkin, 1992, 1997), or were taking the course as an elective

(Aleamoni, 1981; Braskamp & Ory, 1994; Centra, 1993; Feldman, 1978). The IDEA motivation item, "I really wanted to take this course regardless of who taught it," correlates positively with overall excellence of the teacher (r = .22), overall excellence of the course (r = .50), and student ratings of progress on all 12 learning objectives (average r = .28) (Hoyt & Lee, 2002a). This motivation item is one variable that is used to adjust IDEA student ratings for the influence of extraneous factors beyond the instructor's control (see Hoyt & Lee, 2002a, pp. 36-43).

Marsh (2007) also concluded that the reason for taking a course (which overlaps with student motivation) is related to student ratings. This variable is *not a bias*, because motivated students are likely to learn more. However, because motivation to take the course is a student characteristic, and *not* necessarily a reflection of the instructor's teaching effectiveness, this variable REQUIRES SOME CONTROL.

Possibly related to this, Centra (2009) found that required courses tended to receive lower ratings than other kinds of courses, but the differences were not great. Nonetheless, it would *not* be fair to penalize instructors teaching required courses or appropriate to reward those teaching an elective course. Expressing another perspective, Hoyt and Cashin (1977) found that some "required" courses were very popular with students (especially required courses in the major), and some "elective" courses were regarded less positively (especially science or mathematics electives taken to satisfy distribution requirements). Measures of motivation/interest in the course have therefore been shown to be more useful as a control variable.

2. **Expected grades.** In a study involving over 50,000 classes, Centra (2003) examined the relationship between expected grades and student ratings of the contribution of the quality of instruction to their learning. Controlling for class size, teaching method, and student ratings of progress on learning outcomes, expected grade generally had no effect on ratings across eight subject matter areas. However, others have reported positive but low correlations (.10 to .30) between student ratings and expected grades (Braskamp & Ory, 1994; Centra, 2003; Feldman, 1976a, 1997; Howard & Maxwell, 1980, 1982; Marsh & Dunkin, 1992, 1997; Marsh & Roche, 2000).

   Three possible hypotheses have been proposed for these low positive correlations. The *validity hypothesis* posits students who learn more earn higher grades *and* give higher ratings (which supports the validity of student ratings). Another explanation is *grading leniency*: instructors who give higher grades than the students deserve receive higher ratings than the instructors deserve. A third hypothesis is that *student characteristics* (e.g., high interest or motivation) lead to greater learning and, therefore, higher grades *and* higher ratings. Related to this, Centra (2003) suggested that students are applying *attribution principles*: they attribute high grades to their hard work and low grades to poor teaching.

   In two studies of IDEA data, Howard and Maxwell (1980; 1982) concluded that most of the correlation between expected grades and global ratings of the instructor was explained by student (self-reported) learning – the validity hypothesis – and desire to take the course – a student characteristic. More recently, Marsh (2007) reviewed studies of the relationship(s) between student ratings and expected grades. In general, the results supported the validity hypotheses, with some support for student characteristics.

   McKeachie's (1979) explanation for the correlation between grades and ratings still seems appropriate: "[I]n courses in which students learn more the grades should be higher and the ratings should be higher so that a correlation between average grades and ratings is not necessarily a sign of invalidity" (p. 391). To control for the possibility of grading leniency, however, one might have peers (faculty knowledgeable in the subject matter) review the course material, especially exams, test results, graded samples of essays, projects, and so forth to judge the course standards and the bases for grading in the course (McKeachie, 1979).

C. Course variables *related* to student ratings:

1. **Level of the course.** Although we reported previously that *level of the student* is unrelated to student ratings, higher-level courses (especially graduate courses) are rated somewhat higher than lower-level courses (Aleamoni, 1981; Braskamp & Ory, 1994; Feldman, 1978). However, the differences tend to be small.

   Regarding possible control, institutions should check to see if their lower-level classes receive lower ratings than their upper-level classes; similarly they should compare undergraduate with graduate classes. If differences exist, do they remain after controlling for student motivation and class size? If so, we recommend developing local comparative data for the appropriate levels.

2. **Class size.** Although there is a tendency for smaller classes to receive higher ratings, it is a very weak inverse relationship (average r = -.09) (Feldman, 1984). Hoyt and Lee (2002a) found that the effect of class size on ratings was not always statistically significant, but when it was the relationship was negative. Instructors teaching small classes therefore have a slight advantage over those teaching large classes. Consequently, in the interest of fairness, scores on the individual IDEA class report are adjusted for class enrollment.

Centra (2009) found that smaller classes not only tend to receive higher ratings but that students in those classes report learning more. Thus, class size is related to both student learning and effective teaching. Consequently, class size is *not* considered a bias. However, Centra suggested that institutions might want to take class size into consideration – by using comparative data – when considering student ratings in personnel decisions.

3. **Academic discipline.** Feldman (1978) reviewed studies showing that humanities and arts courses receive higher ratings than social science courses, which in turn receive higher ratings than math and science courses. Others (Braskamp & Ory, 1994; Cashin, 1990; Centra, 1993, 2009; Hoyt & Lee, 2002b; Marsh & Dunkin, 1992; Sixbury & Cashin, 1995) found similar results. Although there is increasing evidence that ratings differ between disciplines, it is *not clear why*. Cashin (1990) suggested some possible explanations. For example, some fields may be rated lower because they are more poorly taught; if so, then these differences do *not* require control. However, if instructors in fields requiring more quantitative reasoning skills are rated lower because today's students are less competent in such skills – one of the hypotheses offered to explain why some disciplines are rated lower – then some control is necessary. Centra (2009), in fact, found that mathematical/science courses do tend to receive lower ratings. He suggested that institutions might want to use comparative data to determine if the lower ratings may be the result of lower student quantitative skills.

   With respect to IDEA student ratings, Hoyt and Lee (2002b) examined differences across 28 academic disciplines. For comparisons of learning objectives, teaching methods, student and course characteristics, and global outcome measures by discipline, see IDEA Technical Report No. 13 (Hoyt & Lee, 2002b).

4. **Workload/difficulty.** Course workload and subject-matter difficulty are correlated with student ratings (Centra, 1993, 2003; Marsh, 2001; Marsh & Roche, 2000). Contrary to what some might believe, the correlations are positive – students give somewhat *higher* ratings to difficult courses that require hard work. Still, the correlations are not large. Greenwald and Gillmore (1997) reported just the opposite – that courses with lighter workloads received *higher* student ratings. However, Marsh (2001) re-analyzed their data and found two nearly *un*correlated components of workload: "bad workload" (time spent that was *not* valuable) and "good workload" (i.e., time spent on activities related to instructional objectives). Whereas "bad workload" was correlated negatively with student ratings, "good workload" (work that helps students learn) was positively correlated.

   Hoyt and Lee (2002a) controlled for the instructor's influence (amount of reading, amount of other work, stimulating students' intellectual effort) on student perceptions of the difficulty of the subject matter. They computed a residual score that represented the students' perception of difficulty once the instructor's influence had been removed. If students' perceived the discipline as difficult, ratings were usually somewhat lower. However, difficulty was *positively* correlated with student progress on basic cognitive objectives related to factual knowledge and learning of principles and theories.

   A few researchers (Centra, 2003; Marsh & Roche, 2000; Marsh, 2001) have reported a non-linear relationship between workload/difficulty and student ratings. For example, Centra (2003), using a large database of classes, found that courses were rated lower when they were perceived as either too difficult or too elementary; the highest evaluations were found in classes where difficulty/workload was rated as "just right." However, the relationship was not strong.

   Because of the relationship between workload/difficulty and student ratings, IDEA controls for these variables in its adjusted scores. See a description of this process in Hoyt and Lee (2002a, pp. 36-39). Student ratings of the "difficulty of subject matter" (Item 35) are used to adjust ratings after controlling for the instructor's influence on course difficulty. Students' judgments of how much effort, in general, they put forth on academic work are also used to adjust the scores (Item 43).

   To sum up this section, relatively few variables are related to student ratings that are not also correlated with instructional effectiveness. Nonetheless, a few student and course variables may require some control. In the following paragraphs, we address administration procedures that can affect student ratings when not controlled.

Validity Approach Four: Manipulating Administrative Procedures

Non-anonymous ratings. Students tend to give higher course and instructor ratings when they surrender their anonymity by signing the ratings (Braskamp & Ory, 1994; Centra, 1993; Feldman, 1979; Marsh & Dunkin, 1992). Requiring students to sign their names may inflate the ratings because some students may be concerned about possible reprisals. Suggested control: instructors should urge students *not* to sign their ratings.

Instructor present while students complete ratings. Ratings tend to be higher (Braskamp & Ory, 1994; Centra, 1993; Feldman, 1979; Marsh & Dunkin, 1992) when the instructor is present, possibly for the same reason as non-anonymous ratings. Suggested control: the instructor should leave the room, and a neutral person should collect the ratings.

Purpose of the ratings. Some researchers have investigated whether the directions given to students affect ratings. Centra (1976) found ratings of the instructor's overall effectiveness did not differ between conditions specifying that ratings would be used for *personnel decisions* versus used only by the instructor for improvement. In reviewing Centra's (1976) results, Feldman (1979) further noted that the effect of instructions on ratings varied by the teacher. In some cases, specifying that the ratings would be used for tenure, salary, and promotion decisions resulted in higher ratings, whereas in others it had no effect or was associated with lower ratings. So, the effect of varying the directions on student ratings is small (Marsh, 2007) and inconsistent. Suggested control: instructors should include in the standard directions the intended purpose(s) of the ratings. Although this will *not* eliminate potential bias, it will control *variations* in ratings due to differences in student beliefs about how they will be used.

Validity Approach Five: Analyzing the Underlying Dimensions of Ratings

There is broad agreement that student ratings are multidimensional (i.e., that they reflect several different aspects of teaching). The number of dimensions varies depending, in part, on the form studied and the number and kind of individual items it contains. Put simply, multidimensionality suggests *no single student ratings item or set of related items is useful for all purposes*. There have been a number of factor-analytic studies (see Abrami & d'Apollonia, 1990; Hoyt & Lee, 2002a; Kulik & McKeachie, 1975; Marsh & Dunkin, 1992) in which the dimensions were derived statistically. In several of his reviews of the literature, Feldman (1976b, 1983, 1984, 1987, and 1988) categorized student ratings items (and gave examples) into as many as 22 different logical dimensions. In a later review, Feldman (1989b, 2007) identified 28 dimensions.

Both Centra (1993) and Braskamp and Ory (1994) identified six factors commonly found in student-rating forms:

1. course organization and planning;
2. clarity, communication skills;
3. teacher student interaction, rapport;
4. course difficulty, workload;
5. grading and examinations; and
6. student self-rated learning.

Marsh's (1984, 2007) *Students' Evaluations of Educational Quality* (SEEQ) form has nine dimensions: learning/value, enthusiasm, organization, group interaction, individual rapport, breadth of coverage, exams/grades, assignments, and workload. Other student-rating instruments have items measuring some or all of the above dimensions. Hoyt and Lee (2002a) reported five dimensions of teaching based on IDEA *Diagnostic Form* Items 1 to 20: 1) providing a clear classroom structure, 2) stimulating student interest, 3) stimulating student effort, 4) involving students, and 5) student interaction.

The consistent multidimensionality found in ratings suggests students can distinguish among factors related to teaching effectiveness. Moreover, students can differentially weight teaching behaviors when making overall evaluations of the instructor. Ryan and Harrison (1995), for example, found that amount learned and exam fairness were the two most important criteria students used in making judgments about an instructor's performance.

When using student ratings data to improve teaching, instructors should distinguish among the various items and their factor structure to insure that all of the appropriate dimensions of teaching are rated. Hoyt and Lee (2002a) found that the relevance of 20 different IDEA teaching methods varied depending upon which learning objectives were emphasized in a course. The implication was that different kinds of learning may require different types of teaching.

Although there is general agreement that student ratings are multidimensional, and that various dimensions should be used when their purpose is to improve teaching, there is disagreement about how many and which dimensions should be used for personnel decisions (Apodaca & Grad, 2005; Harrison, Douglas, & Burdsal, 2004; Hobson & Talbot, 2001; Renaud & Murray, 2005). In several articles, Abrami (e.g., Abrami & d'Apollonia, 1991) suggested that one or a few global/summary items might be sufficient for personnel decisions. Others have made a similar recommendation (e.g., Braskamp & Ory, 1994; Cashin & Downey, 1992; and Centra, 1993). Harrison and colleagues also confirmed that various weighted and un-weighted measures of overall evaluations of teaching effectiveness are highly inter-correlated (Harrison, Douglas, & Burdsall, 2004).

Offering a relevant view, McKeachie (1997) argued that, when it comes to personnel decisions, student ratings of attainment of educational goals and objectives are preferable to multiple dimensions or a single measure of overall teaching effectiveness. Effective teaching can be demonstrated in many ways, and no instructor should be expected to demonstrate proficiency in all methods and styles. Moreover, teaching methods may vary, depending upon the course content, student characteristics, and size of class. Regardless of which measures are used, administrators and members of personnel committees should use broad categories (e.g., exceeds expectations, meets expectations, fails to meet expectations) rather than try to interpret decimal point differences (d'Apollonia & Abrami, 1997; McKeachie, 1997; Pallett, 2006).

The research cited thus far has summarized evidence of the validity of student ratings as found in correlations with student achievement, correlations with other

criteria, examinations of potential bias, manipulations of administration procedures, and factor-analytic studies. In the next sections, we summarize findings from investigations of online ratings and the usefulness of student ratings.

Student Ratings Administered Online versus on Paper

The use of online student ratings has increased steadily with the growth of Web-based surveys. Online delivery offers several advantages over paper-and-pencil administration. Students can respond outside of class at their convenience, freeing class time for other activities (Dommeyer, Baum, & Hanna, 2003; Layne, DeCristoforo, & McGinty, 1999). Response rates to open-ended questions posted online tend to be higher (Johnson, 2003) and written comments lengthier (Hardy, 2003; Johnson, 2003; Layne et al., 1999). Moreover, online directions and procedures can be uniform for all classes, enabling instructors to be less involved in the administration process (Layne et al., 1999).

The chief disadvantage of online ratings is lower student response rates to the fixed items, which threaten class representativeness (Sorenson & Reiner, 2003). Lower response rates occur for several reasons, among them student concern about anonymity, computer technical difficulties, and the time required to respond outside of class (Dommeyer et al., 2003). Some instructors fear lower response rates create a negative bias because students who are dissatisfied with the course or instructor might be more likely than others to respond (Johnson, 2003). However, correlations between response rate and overall ratings of the instructor and course are, on average, quite low (Benton, Webster, Gross, & Pallett, 2010a; Johnson, 2003), which suggests response bias is less likely.

In spite of the disparity in response rates, researchers have consistently found no meaningful differences between online and paper surveys. When the same students respond under both formats, the correlations are high between global ratings of the instructor (.84) and course (.86) (Johnson, 2003). Further, no meaningful differences are found in individual item means, number of positive and negative written comments (Venette, Sellnow, & McIntire, 2010), scale means and reliabilities, and the underlying factor structure of the ratings (Leung & Kember, 2005). Similarly, when different students respond to online and paper surveys, no meaningful differences are found in student progress on relevant course objectives, global ratings of the course and instructor, frequency of various teaching methods (Benton et al., 2010a), subscale means (Layne et al., 1999), the proportion of positive and negative written comments (Hardy, 2003), and the underlying factor structure (Layne et al., 1999).

Suggestions for increasing online response rates: Higher online response rates are more likely when instructors clearly communicate their expectations for compliance. Online response rates also tend to be higher when students complete ratings for more than one course (Johnson, 2003). Ensuring student confidentiality, monitoring response rates, encouraging instructor follow-up, sending reminders, acknowledging and rewarding high response rates, and integrating the process into the campus culture may also be associated with higher response rates (see The IDEA Center, 2008).

Student Ratings in Face-to-Face versus Online Courses

Because of the differences between online and face-to-face classroom environments, some have investigated whether instructors can use the same student ratings approach in online courses (e.g., Beattie, Spooner, Jordan, Algozzine, & Spooner, 2002; Benton, Webster, Gross, & Pallett, 2010b). The online environment may either diminish or enhance opportunities for student participation (for example, via online posts), hinder or facilitate expression of ideas (especially for those who are reluctant to speak in class), reduce or increase access to the instructor (via e-mail), and moderate or expand connections with other students (via "chat rooms"), all of which might affect student ratings positively or negatively (Smith, Smith, & Boone, 2000).

However, student ratings collected in face-to-face and online courses are actually more similar than they are different. Student progress on relevant objectives, global ratings of the course and instructor, and the frequency of various teaching methods are comparable between courses identified exclusively as either face to face or online (Benton et al., 2010b). Moreover, individual item means, internal consistency reliabilities, and the underlying factor structures are very similar between ratings collected online from students enrolled in distance courses and ratings collected on paper from students enrolled on campus (McGhee & Lowell, 2003). Furthermore, item means and the overall assessment of the instructor are nearly identical between students enrolled in multiple online and face-to-face sections of the same course taught by the same instructor (Wang & Newlin, 2000).

Nonetheless, some differences do exist. As one might expect, response rates are somewhat lower in online courses. However, the correlations between response rate and overall ratings of the instructor and course are, on average, low, making negative response bias to online ratings less likely (Benton et al., 2010b). Not surprisingly, students in online courses report greater instructor use of educational technology to promote learning, and such use is more highly correlated with student progress in online courses. In addition, students report somewhat more reading in online courses (Benton et al., 2010b).

Usefulness of Student Ratings

Cohen (1980) performed a meta-analysis of 17 studies on the effect of student-ratings feedback on improving teaching. Receiving feedback about student ratings administered during the first half of the term was *positively*

related to improving college teaching as measured by student ratings administered at the end of the term. In the typical study, there were three groups. All groups had ratings administered during the first half of the semester and again at the end. The first group received no feedback. The second group received student-ratings feedback provided in the quantitative data from the first student ratings. The third group received student-ratings feedback and, in addition, some kind of consultation (the quality of which varied across studies). Cohen used the end-of-term ratings as the measure of improvement and set the first group's mean ratings at the 50th percentile (see Figure 4). As indicated in Figure 4, if an institution really intends to use student ratings to improve teaching, some kind of consultation for instructors is recommended.

Figure 4 • Effect of Student Rating Feedback on Improving End-of-term Ratings of Teaching.

Treatment Group	End of Term Percentile Rank
No student rating feedback	50th
Only student rating feedback	58th
Student rating feedback plus consultation	74th

Additional research indicates that combining consultation with feedback is significantly more useful for improving teaching than feedback alone (Brinko, 1990; Hampton & Reiser, 2004; Marincovich, 1999). Discussing ratings with a peer or consultant improves their usefulness (Aleamoni, 1978; Marsh & Overall, 1979). Feedback and consultation that target problems identified by students and that address specific teaching behaviors results in the greatest improvement (Marsh & Roche, 1993). Faculty find feedback about interaction with students especially helpful, followed by grading practices, global ratings of the course and instructor, and structural issues (e.g., pace of course, exam difficulty and content, and textbook) (Schmelkin, Spencer, & Gellman, 1997).

Penny and Coe (2004) conducted a meta-analysis of 11 studies of the effectiveness of consultation on student ratings feedback. They concluded that the following eight strategies led to the most effective consultation for improvement (p. 245):

1. active involvement of teachers in learning process;
2. use of multiple sources of information (e.g., videotapes);
3. interaction with peers;
4. sufficient time for dialogue and interaction;
5. use of teacher self-ratings;
6. use of high-quality feedback information (e.g., student ratings);
7. examination of conception of teaching; and
8. setting of improvement goals.

In the absence of a consultant, instructors should reflect on what the ratings mean as a useful first step. Kember and colleagues developed a four-category scheme for assessing quality of self-reflection (Kember et al., 2008). In *nonreflection*, an instructor simply looks through the ratings without giving them much thought. At the second level of *understanding*, the instructor attempts to grasp what the ratings mean but does not relate them to his or her own experiences. It is not until *reflection* that instructors relate the results to their own experience teaching the specific course. Finally, in *critical reflection* the teacher undergoes a transformation in perspective, perhaps brought on by the disequilibrium or cognitive dissonance produced when the feedback from student ratings differs from the teacher's view of how things went.

Such feedback can be humbling, but it may lead instructors to admit that something in the course or their teaching needs to change (Weimer, 2009). Meaningful change, according to instructors who have made significant improvements in end-of-course ratings, does not require great effort (McGowan & Graham, 2009). Improvements in ratings are most frequently associated with creating opportunities for active learning in the classroom, fostering better student-teacher interactions, setting expectations and maintaining high standards, being prepared for class, and revising procedures for assessing student work (McGowan & Graham, 2009).

Unfortunately, the actual use of student ratings for formative purposes falls far short of its potential. Pallett (2006) suggested three reasons for this shortcoming. First, institutions sometimes place too much emphasis on the summative component of ratings. When student ratings are overemphasized for summative evaluation and underutilized for developmental purposes, faculty often lose trust in the process and see little or no benefit in collecting student feedback. Such misuse erodes the potential benefits of ratings and can create a negative climate for faculty evaluation. A second reason student ratings tend to be underutilized for formative purposes is the difficulty associated with creating valid and reliable ratings instruments that provide helpful feedback. Third, at some institutions there is insufficient mentoring. Credible mentors who are trusted colleagues, not necessarily involved in personnel decisions, should be available to provide feedback and make recommendations for improvement.

Conclusion

There are probably more studies of student ratings than of all of the other data used to evaluate college teaching combined. Although one can find individual studies that support almost any conclusion, for many variables there are enough studies to discern trends. In general, student ratings tend to be statistically reliable, valid, and relatively free from bias or the need for control, perhaps more so than any other data used for faculty evaluation.

Nonetheless, student ratings are *only one source* of data about teaching and must be used in combination with multiple sources of information if one wishes to make a judgment about all of the components of college teaching. Further, student ratings must be interpreted. We should not confuse a source of data with the evaluators who use it – in combination with other kinds of information – to make judgments about an instructor's teaching effectiveness (Cashin, 2003).

This paper summarizes the general conclusions from the research on student ratings. Whether these conclusions hold true for all contexts is an empirical question. If an institution has reason to believe that these conclusions do *not* apply, key players should gather local data to address the issue. In the absence of evidence to the contrary, however, the following general conclusions can be used as a guide (Marsh, 2007, p. 372):

> SETs [student evaluations of teaching effectiveness] are multidimensional, reliable and stable, primarily a function of the instructor who teaches a course rather than the course that is taught, relatively valid against a variety of indicators of effective teaching, relatively unaffected by a variety of potential biases, and are seen to be useful by faculty, students, and administrators.

References

Abrami, P. C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them? New Directions for Institutional Research*, No. 109 (pp. 59-87). San Francisco: Jossey-Bass.

Abrami, P. C., & d'Apollonia. (1990). The dimensionality of ratings and their use in personnel decisions. In M. Theall & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice: New Directions for Teaching and Learning*, No. 43 (pp. 97-111). San Francisco: Jossey-Bass.

Abrami, P. C., & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness- generalizability of "N = 1" research: Comments on Marsh (1991). *Journal of Educational Psychology*, 83, 411-415.

Abrami, P. C., d'Apollonia, S., & Rosenfeld, S. (2007). The dimensionality of student rations of instruction: An update on what we know, do not know, and need to do. In R. P. Perry & J. C. Smart (Eds.), *The Scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 385-445). Dordrecht, The Netherlands: Springer.

Abrami, P. C., Leventhal, L., & Perry, R. P. (1982). Educational seduction. *Review of Educational Research*, 52, 446-464.

Abrami, P. C., Perry, R. P., & Leventhal, L. (1982). The relationship between student personality characteristics, teacher ratings, and student achievement. *Journal of Educational Psychology*, 74, 111-125.

Abrami, P. C., Rosenfeld, S., & Dedic, H. (2007). Commentary: The dimensionality of student rations of instruction: What we know, and what we do not. In R. P. Perry & J. C. Smart (Eds.), *The Scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 385-446). Dordrecht, The Netherlands: Springer.

Aleamoni, L. M. (1981). Student ratings of instruction. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 110-145). Beverly Hills, CA: Sage.

Aleamoni, L. M. (1978). The usefulness of student evaluations in improving college teaching. *Instructional Science*, 7, 95-105.

Aleamoni, L. M. (1987). Typical faculty concerns about student evaluation of teaching. In L. M. Aleamoni (Ed.), *Techniques for evaluating and improving instruction: New Directions for Teaching and Learning*, No. 31 (pp. 25-31). San Francisco: Jossey-Bass.

Apodaca, P. & Grad, H. (2005). The dimensionality of student ratings of teaching: Integration of uni- and multidimensional models. *Studies in Higher Education*, 30, 723-748.

'd Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist*, 52, 1198-1208.

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

Beattie, J., Spooner, F., Jordan, L., Algozzine, B., & Spooner, M. (2002). Evaluating instruction in distance learning classes. T*eacher Education and Special Education*, 25, 124-132.

Benton, S. L., Duchon, D., & Pallett, W. H., (2011). Validity of student self-reported ratings of instruction. *Assessment & Evaluation in Higher Education*. http://dx.doi.org/10.1080/02602938.2011.636799.

Benton, S. L., Webster, R., Gross, A., & Pallett, W. (2010a). *IDEA Technical Report No. 16: An analysis of IDEA Student Ratings of instruction using paper versus online survey methods*. Manhattan, KS: The IDEA Center.

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

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

Braskamp, L. A., Ory, J. C., & Pieper, D. M. (1981). Student written comments: Dimensions of instructional quality. *Journal of Educational Psychology*, 73, 65-70.

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

Burdsal, C. A., & Harrison, P. D. (2008). Further evidence supporting the validity of both a multidimensional profile and an overall evaluation of teaching effectiveness. *Assessment and Evaluation in Higher Education*, 33, 567-576.

Carrier, N. A., Howard, G. S., & Miller, W. G. (1974). Course evaluations: When? *Journal of Educational Psychology*, 66, 609-613.

Cashin, W. E. (1989). *Defining and evaluating college teaching*. IDEA Paper No. 21. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E. (1990). Students do rate different academic fields differently. In M. Theall, & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice: New Directions for Teaching and Learning*, No. 43 (pp. 113-121). San Francisco: Jossey-Bass.

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited*. IDEA Paper No. 32. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E. (2003). Evaluating college and university teaching: Reflections of a practitioner. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 531-593). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

Cashin, W. E., Downey, R. G., & Sixbury, G. R. (1994). Global and specific ratings of teaching effectiveness and their relation to course objectives: Reply to Marsh (1994). *Journal of Educational Psychology*, 86, 649-657.

Centra, J. A. (1976). The influence of different directions on student ratings of instruction. *Journal of Educational Measurement*, 13, 277-282.

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco: Jossey-Bass.

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. (2009). *Differences in responses to the Student Instructional Report: Is it bias?* Princeton, NJ: Educational Testing Service.

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

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

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.

Cohen, P. A. (1987, April). *A critical analysis and reanalysis of the multisection validity meta-analysis*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Costin, F. (1968). A graduate course in the teaching of psychology: Description and evaluation. *Journal of Teacher Education*, 19, 425-432.

Davis, B. G. (2009). *Tools for teaching*, (2nd ed.). San Francisco: Jossey-Bass.

Dommeyer, C. J., Baum, P., & Hanna, R. W. (2003). College students' attitudes toward methods of collecting teaching evaluation: In-class versus on-line (Electronic version). *Journal of Education for Business*, 78, 5-11.

Erdle, S., Murray, H. G., & Rushton, J. P. (1985). Personality, classroom behavior, and student ratings of college teaching effectiveness: A Path Analysis. *Journal of Educational Psychology*, 77, 394-407.

Feldman, K. A. (1976a). Grades and college students' evaluations of their courses and teachers. *Research in Higher Education*, 4, 69-111.

Feldman, K. A. (1976b). The superior college teacher from the students' view. *Research in Higher Education*, 5, 243-288.

Feldman, K. A. (1977). Consistency and variability among college students in rating their teachers and courses: A review and analysis. *Research in Higher Education*, 6, 233-274.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Educatio*n, 9, 199-242.

Feldman, K. A. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education*, 10, 149-172.

Feldman, K. A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. *Research in Higher Education*, 18, 3-124.

Feldman, K. A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education*, 21, 45-116.

Feldman, K. A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis. *Research in Higher Education*, 24, 129-213.

Feldman, K. A. (1987). Research productivity and scholarly accomplishment of college teachers as related to their instructional effectiveness: A review and exploration. *Research in Higher Education*, 26, 227-298.

Feldman, K. A. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities. *Research in Higher Education*, 28, 291-344.

Feldman, K. A. (1989a). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators and external (neutral) observers. *Research in Higher Education*, 30, 137-194.

Feldman, K. A. (1989b). 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.

Feldman, K. A. (1992). College students' views of male and female college teachers: Part I-Evidence from the social laboratory and experiments. *Research in Higher Education*, 33, 317-375.

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.

Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practic*e (pp. 368-395). New York: Agathon Press.

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *The Scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 93-129). Dordrecht, The Netherlands: Springer.

Forsyth, D. R. (2003). *Professor's guide to teaching: Psychological principles and practices*. Washington, DC: American Psychological Association.

Frey, P. W. (1976). Validity of student instructional ratings as a function of their timing. *Journal of Higher Education*, 47, 327-336.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement*, 15, 1-13.

Greenwald, A. G. & Gillmore, G. M. (1997). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology*, 89, 743-751.

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.

Hardy, N. (2003). Online ratings: Fact and fiction. In T. D. Johnson & D. L. Sorenson (Eds.), *Online student ratings of instruction: New Directions for Teaching and Learning*, No. 96 (pp. 31-38). San Francisco: Jossey-Bass.

Harrison, P. D., Douglas, D. K., & Burdsall, C. A. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education*, 45, 311-323.

Hobson, S. M., & Talbot, D. M. (2001). Understanding student evaluations: What all faculty should know. *College Teaching*, 49, 26-31.

Hogan, T. P. (1973). Similarity of student ratings across instructors, courses, and time. *Research in Higher Education*, 1, 149-154.

Hornbeak, J. L. (2009). Teaching methods and course characteristics related to college students' desire to take a course. Kansas State University. *Dissertations and Theses: Full Text*. AAT 3358789.

Howard, G. S., & Maxwell, S. E. (1980). The correlation between student satisfaction and grades: A case of mistaken causation? *Journal of Educational Psychology*, 72, 810-820.

Howard, G. S., & Maxwell, S. E. (1982). Do grades contaminate student evaluations of instruction? *Research in Higher Education*, 16, 175-188.

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

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

Hoyt, D. P., & Lee, E. (2002b). *Technical Report No. 13: Disciplinary differences in student ratings*. Manhattan, KS: The IDEA Center.

IDEA Research Note 1. (2003). *The "excellent teacher" item*. Manhattan, KS: The IDEA Center.

Johnson, T. D. (2003). Online student ratings: Will students respond? In T. D. Johnson & D. L. Sorenson (Eds.), *Online student ratings of instruction: New Directions for Teaching and Learning*, No. 96 (pp. 49-59). San Francisco: Jossey-Bass.

Kember, D., McKay, J., Sinclair, K. & Wong, F. K. Y. (2008). A four-category scheme for coding and assessing the level of reflection in written work. *Assessment & Evaluation in Higher Education*, 33(4), 363-379.

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them? New Directions for Institutional Research*, No. 109 (pp. 9-25). San Francisco: Jossey-Bass.

Kulik, J. A., & McKeachie, W. J. (1975). The evaluation of teachers in higher education. In F. N. Kerlinger (Ed.), *Review of research in education* (Vol. 3, pp. 210-240). Itasca, IL: F. E. Peacock.

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

Leung, D. Y. P., & Kember, D. (2005). Comparability of data gathered from evaluation questionnaires on paper through the Internet. *Research in Higher Education*, 46, 571-591.

Li, Y. (1993). A comparative study of Asian and American students' perceptions of faculty teaching effectiveness at Ohio University. *Dissertations & Theses: Full Text*, Ohio University, Athens. AAT 9335076.

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. (1982). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology*, 74, 264-279.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, 76, 707-754.

Marsh, H. W. (2001). Distinguishing between good (useful) and bad workloads on student evaluations of teaching. *American Educational Research Journal*, 38, 183-212.

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., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research*, Vol. 8. New York: Agathon Press.

Marsh, H. W., & Dunkin, M. J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241-320). New York: Agathon Press.

Marsh, H. W., & Hattie, J. (2002). The relation between research productivity and teaching effectiveness. *Journal of Higher Education*, 73, 603-641.

Marsh, H. W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching & Teacher Education*, 7, 303-314.

Marsh, H. W., & Overall, J. U. (1979). Long-term stability of students' evaluations: A note on Feldman's consistency and variability among college students in rating their teachers and courses. *Research in Higher Education*, 10, 139-147.

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluation by their students. *Journal of Educational Psychology*, 71, 149-160.

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.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, and innocent bystanders. *Journal of Educational Psychology*, 92, 202-22.

Marsh, H. W., & Ware, J. E. (1982). Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretations of the Dr. Fox effect. *Journal of Educational Psychology*, 74, 126-134.

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

McGowan, W. R., & Graham, C. R. (2009). Factors contributing to improved teaching performance. *Innovative Higher Education*, 34, 161-171.

McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe*, 65, 384-397.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist*, 52, 1218-1225.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). Old Tappan, NJ: Macmillan, 1989.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 1995, 50, 741–749.

Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology*, 75, 138-149.

Murray, H. G. (2007). Low-inference teaching behaviors and college teaching effectiveness: Recent developments and controversies. In R. P. Perry & J. C. Smart (Eds.), *The Scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 145-200). Dordrecht, The Netherlands: Springer.

Murray, H. G., Rushton, J. P., & Paunonen, S. V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology*, 82, 250-261.

Naftulin, D. H., Ware, J. E., & Donnelly, F. A. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education*, 48, 630-635.

Ory, J. C., Braskamp, L. A., & Pieper, D. M. (1980). Congruency of student evaluative information collected by three methods. *Journal of Educational Psychology*, 72, 181-185.

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? In T. D. Johnson & D. L. Sorenson (Eds.), *Online student ratings of instruction: New Directions for Teaching and Learning*, No. 5 (pp. 27-44). San Francisco: Jossey-Bass.

Overall, J. U., & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology*, 72, 321-325.

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

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

Perry, R. P., & Smart, J. C. (Eds.) (1997). *Effective teaching in higher education: Research and practice*. New York: Agathon Press.

Perry, R. P., & Smart, J. C. (Eds.) (2007). *The Scholarship of teaching and learning in higher education: An evidence-based perspective*. Dordrecht, The Netherlands: Springer.

Renaud, R. D., & Murray, H. G. (1996). Aging, personality, and teaching effectiveness in academic psychologists. *Research in Higher Education*, 37, 323-340.

Renaud, R. D., & Murray, H. G. (2005). Factorial validity of student ratings of instruction. *Research in Higher Education*, 46, 929-953.

Ryan, J. M., & Harrison, P. D. (1995). The relationship between individual instructional characteristics and the overall assessment of teaching effectiveness across different contexts. *Research in Higher Education*, 34, 213-228.

Schmelkin, L. P., Spencer, K. J., & Gellman, E. S. (1997). Faculty perspectives on course and teacher evaluations. *Research in Higher Education*, 38, 575-592.

Sixbury, G. R., & Cashin, W. E. (1995). *IDEA technical report no. 10: Comparative data by academic field*. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Smith, S. B., Smith, S. J., & Boone, R. (2000). Increasing access to teacher preparation: The effectiveness of traditional instructional methods in an online learning environment. *Journal of Special Education Technology*, 15(2), 37-46.

Sorenson, D. L., & Reiner, C. (2003). Charting the uncharted seas of online student ratings of instruction. In T. D. Johnson & D. L. Sorenson (Eds.), *Online student ratings of instruction: New Directions for Teaching and Learning*, No. 96 (pp. 1-24). San Francisco: Jossey-Bass.

Svinicki, M., & McKeachie, W. J. (2011). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers* (13th ed.). Belmont, CA: Wadsworth.

Theall, M., & Feldman, K. A. (2007). Commentary and update on Feldman's (1997) "Identifying exemplary teachers and teaching: Evidence from student ratings." In R. P. Perry & J. C. Smart (Eds.), *The teaching and learning in higher education: An evidence-based perspective* (pp. 130-143). Dordrecht, The Netherlands: Springer.

The IDEA Center (2008). Facilitating response rates in IDEA Online. Retrieved from http://www.theideacenter.org/OnlineResponseRates. Manhattan, KS: The IDEA Center.

Venette, S., Sellnow, D., & McIntire, K. (2010). Charting new territory: Assessing the online frontier of student ratings of instruction. *Assessment & Evaluation in Higher Education*, 35, 101-115.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, 23, 191-211.

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

Weimer, M. (2009). Teachers who improved. *The Teaching Professor*, 23, 2.

T: 800.255.2757
T: 785.320.2400
F: 785.320.2424

E: info@theideacenter.org
www.theideacenter.org