ETS SIR II

Listening. Learning. Leading.®

The Student Instructional Report II™: Its Development, Uses and Supporting Research

An Issue Paper from ETS

www.ets.org

Introduction

The concept of students "grading" their instructors — or at least evaluating the quality of the instruction they receive — arrived on college campuses in the mid-1960s. The number of colleges offering that option to students has been on the rise ever since.

Today, there is hardly a college or university that does not conduct some student rating of instruction programs, and the single most widely used evaluation instrument over the years has been the ETS® Student Instructional Report (SIR). It was introduced in 1972 and updated by ETS in the mid-1990s as the SIR IITM report. Millions of students in tens of thousands of classrooms have used SIR or SIR II reports.

In this paper, I will discuss the development of the SIR II report; describe research that has been conducted with both forms of the instrument to address important questions of validity and reliability; and provide some thoughts on how the research results can best be interpreted and used by instructors and colleges. I will also refer to ETS's eSIR report, an adaptation of the SIR II report used to evaluate distance learning courses.

SIR and SIR II reports have been the basis of a dozen research studies. Most of these have been published in professional journals and have appeared in SIR Research Reports published by ETS. Study results are applicable to either instrument because of their similar items and factor structures. Some of the research findings can also be applied to the eSIR, although specific studies with eSIR report are planned.

The Development of SIR II

The original SIR report was based on what was known in 1972 about effective college teaching and how students might contribute to it by evaluating the quality of instruction. Various constituents (teachers, students, administrators) identified characteristics of effective teaching and these characteristics formed the basis for SIR report item selection.

But our knowledge of teaching and evaluating teaching quality increased significantly during the intervening decades. This led to an increased emphasis on assessments that measure what students have learned in courses and what contributions teachers have made toward that learning. It also led to a better understanding of the critical role students' play in learning, as determined by the effort they apply, the time they spend studying and similar factors. As a result, teachers began to use learning tools to complement lectures. A second- generation SIR report had to take these factors into consideration.

The new SIR II report was based on research conducted at ETS on the SIR report, as well as research on college teaching and learning. This led to the development of two new forms, which were created and pre-tested in the spring of 1994. These forms included five scales, on which instruction would be evaluated, from the original instrument with some new items added and old items deleted.

The five scales were:

- Course Organization and Planning
- Communication
- Faculty-Student Interaction
- Assignments, Exams, and Grading
- Course Difficulty and Workload

Three new sets of items were added:

- Course Outcomes
- Student Effort and Involvement
- Supplementary Instructional Methods

These new items reflected the more recent emphasis on outcomes assessments, student effort and the use of learning tools. The two new forms included a different response format to the same set of scales and items. By having students in random halves of 50 classes respond to the two forms, it was possible to analyze the results to determine the better response format.

The original SIR report was analyzed by students and faculty at a single university. ETS conducted the initial testing of the SIR II report at 10 two- and four-year colleges. Traditional item and scale analyses of the two experimental forms included computing means, standard deviations, Coefficient Alphas (reliability estimates), item-to-scale correlations and factor analysis. An additional analysis compared the response formats for the two forms to help determine which provided better variation in student responses. These analyses helped hone and alter the items, and determine the most useful response format.

The response format ultimately selected asked students to rank instructional practices on a five-point scale according to how effective each was in contributing to their learning in the course. The new response scale and the course outcomes scale reflected the new emphasis on the connection between perceived student learning and instructional practices.

The response format rejected was a five-point agree/disagree scale with a neutral midpoint (neither agree nor disagree). Analysis indicated the neutral response merely confounded the results. While several widely used student teaching evaluation instruments currently have a neutral midpoint as an option, it's likely the information they obtain is less useful.

The form selected, the current SIR II report, was pilot tested at a variety of colleges in 1995 and 1996 to investigate its reliability and construct validity. Three kinds of reliability analyses yielded positive results. A Coefficient Alpha analysis indicated that the items were interrelated. An item level reliability analysis examined the extent of

consistency of scores after adjusting group size. When 10 students used the SIR II report, the reliability was marginal (.59); for 15 students it was fair (78); and for 20 students it was very good (.89). This suggested that when the SIR II report is used with small classes, several classes' responses should be evaluated to get a more consistent evaluation. This is particularly important when the results are to be used in personnel decisions. When class responses total over 15 and more than two-thirds of the students in the class have responded, however, a more reliable score can be assumed.

The third kind of reliability, referred to as "test-retest," measures the extent to which responses for a class of students remain stable over short periods of time. When only slight fluctuations occur, we can conclude responses are not subject to the effects of daily occurrences or mood swings, but, in fact, reflect consistent evaluations of instructional characteristics. For this analysis, 42 classes at one college produced uniformly high correlations between the first and second administrations of the SIR II report, indicating that the rank order of student responses did not change much over two weeks

In summary, the pilot testing of the SIR II report demonstrated positive results for the three kinds of reliability.

To provide additional support for the SIR II report's validity, researchers created a set of effective teaching dimensions and test items to reflect them. We were able to obtain a reasonable factor analytic solution for the SIR II report structure that corresponded to these dimensions. Although other factor structures could have resulted in comparable fit to the data, this construct validation approach did indicate that the dimensions SIR II report was intended to measure are a plausible structure for ratings of college-level instruction. Additional studies of validity were later completed with the SIR II report and are discussed later in this paper.

The eSIR for Distance Learning Courses

The eSIR report contains more than half of the items in the SIR II report. A panel of educators involved in distance education suggested the remaining items. The overall Evaluation of Instruction item, which provides important information for teachers and administrators, is identical on both the SIR II and the eSIR reports.

A major feature of the eSIR report is the open-ended comment section at the end of each set of items. Students are encouraged to add or expand on the quantitative ratings anonymously by typing comments. Unlike traditional college courses, which frequently have students from a range of class levels and majors and widely varying levels of motivation, distance learning courses often consist of highly motivated adult learners. Therefore, with the eSIR report, smaller numbers of students may actually produce sufficient reliability. More research will be conducted to examine eSIR report reliability, validity and other issues.

Can the SIR II report lead to Improvement in Instruction?

Significant improvement in instruction after using the SIR II report is likely to take place if four conditions are met:

- Instructors gain new knowledge about their teaching performance and the knowledge could not be gained except by using the SIR II report.
- Instructors accept student evaluations as valuable and useful in improving their instruction.
- The information instructors receive from the evaluations provides helpful advice for how to make changes in instruction.
- The instructors are motivated to make changes. They want to improve either out of a desire to be a better teacher or for job-related reasons like tenure or promotion.

While these conditions may seem obvious, a study I conducted with five colleges that had never used student ratings supports their necessity.

After the SIR report was administered at mid-semester, half of a randomly chosen group of teachers received their results promptly (the feedback experimental group). Half of the teachers did not (the control group). All of the teachers also rated themselves on the items. At the end of the semester the SIR report was administered again to the two groups and to a third, post-test only group. More than 500 teachers participated in the study, including a group that used the SIR report just once a semester later.

Analysis of the results supported my prediction: student ratings would produce changes in teachers who had rated themselves more favorably than their students had rated them. (The self-ratings collected indicated the majority of teachers tend to rate themselves more positively than do students.) It also indicated even more improvements are possible when instructors are provided with additional time and interpretations of the results (i.e., providing comparison data). This indicates some teachers need more time to absorb the feedback from the SIR report and to make changes.

We can conclude that those teachers who made changes and adjusted their instruction valued the new information they received from the student evaluations and used that information to improve their performance. While teachers can adjust their teaching based on student feedback, more significant improvements are likely to occur if faculty development workshops related to SIR II scales are held. For example, one college, mentioned later in this paper, provides workshops on course organization, assignments, exams and other course components measured by the SIR II report.

Do Students Learn More from Teachers They Rate Highly?

This is the validity question frequently asked by faculty members and administrators. They want and need assurance that SIR II results relate to teacher effectiveness and student learning, not to popularity or entertainment value.

A study I completed in the late 1970s found teachers who received higher SIR ratings produced higher achievement test scores for their students. The study was conducted at Memorial University in Newfoundland. The university required all incoming students to take a common liberal arts core of five or six courses and assigned students to sections of each core course at random. Each course had a common final exam made up by teachers in the department who were not teaching a section (making it impossible for instructors to teach to the test). In short, it was an ideal experimental research setting.

The university agreed to administer the SIR report in these classes and provide the ratings and test scores for evaluation. In return they received SIR rating summaries for their classes and the results of the study.

The correlations between ratings and achievement test scores were generally positive, with the single Overall Evaluation item correlating highest. This single item provides a good, global view of instructional effectiveness, unlike other SIR items, which are focused on improving instructional practices. These other items were more modestly correlated with student achievement, because teachers have different strengths and styles of teaching. For example, some teachers are highly organized in their courses and lectures. Others may be less organized, but interact very well with students individually or in small groups. This was the first study to demonstrate the validity of student ratings using an experimental design and an objective criterion (student learning) faculty members believed to be critical to their acceptance of the SIR report as a valid performance measure.

A more recent study conducted with the SIR II report used self-reported student learning (the Course Outcomes scale) to estimate learning and again confirmed SIR II's validity.

Another validity study compared ratings by deans and peers to SIR report results using a teaching portfolio constructed by each faculty member at a college. The SIR evaluations correlated well with the teaching evaluations made by the deans and peers, indicating student views overlap significantly with more experienced evaluators.

Will Teachers Receive Higher SIR II Evaluations by Giving Higher Grades and Less Coursework?

Many faculty members believe they can improve their student evaluations by simply giving students higher grades than they deserve and requiring less work in the course. Research results may never be able to change these perceptions, and it doesn't help that a few studies seem to support their beliefs. Most research findings, however, do not support this position, including a SIR II research study involving more than 50,000 courses.

The large number of courses allowed controls for factors such as class size, teaching method, subject area, and most importantly, student reports of their own learning outcomes. This outcomes measure was important because of the definition of the widely accepted hypothesis at the heart of the study: bias exists when a student, teacher, or course characteristic affects the evaluations made, but is unrelated to any criteria of good teaching.

After controlling for student-reported learning, the study found no correlation between grades students expected to receive and the evaluations they made on SIR II reports. (Expected grade was the appropriate measure because students do not generally know their actual grade in a course when they complete the evaluation form, usually several days or a week before the last class.)

Significantly, teachers of courses rated "just right" in difficulty or workload received the highest ratings. Teachers of courses rated either "too difficult" or "too elementary" received lower ratings. The findings clearly indicate teachers are not likely to improve their SIR II evaluations by simply giving undeserved higher grades or less coursework.

Is There Gender Bias in SIR II Evaluations?

SIR II evaluations from more than 700 different college classes were studied to determine whether male and female students rated the same male and female instructors differently.

The large number of classes eliminated the limitations of previous studies. The sample included 20 different institutions rather than a single college. It analyzed eight academic disciplines as opposed to one. It used the class as a unit of measure, which was more appropriate than measuring individual students. Additionally, classes used in the study were required to have a minimum of 10 female and 10 male students.

Analyses of the mean differences of SIR II evaluations within the same classes and across different classes indicated some same-gender preferences, particularly in female students rating female teachers. But the differences in ratings, though statistically significant, were not large and should not have much impact if the ratings are used in personnel decisions. Moreover, the higher evaluations received by female teachers from female students, and in a few instances by males as well (notably in natural science

courses), could well be due to teaching style. Female instructors who were evaluated in this study were more likely than men to use discussion rather than a lecture method. As a group, the SIR II evaluations also indicated women instructors tended to be more nurturing to students.

Do Students Make Accurate Judgments While Still in Class or in College?

Some teachers believe students can only assess their teaching performance after they have taken later courses or in the later years of their college careers. These teachers believe ratings at the end of a course do not adequately reflect the long-term impact of their instruction.

Studies comparing SIR evaluations of an instructor made by students the following semester, the next year, immediately after graduation and several years later, however, indicate consistent ratings of teachers by students over time.

In one study, SIR ratings of teachers were compared with alumni ratings made five years after graduation. The two sets of ratings were highly correlated. Although students may realize later that a particular subject was more important than they thought, the research suggests student opinions about instructors change very little over time.

Interpreting Scores and SIR II Comparison Data

As with every linear rating scale, responses for the SIR II report do not fall into a normal curve. For example, on five-point SIR II items, a mean class score of 3.6 is numerically above the midpoint. But, when compared to a large sample of classes, it falls below the average or median. To help in interpreting scores, ETS publishes comparative data, which consists of the SIR II scores of past users. Because the sample of institutions does not proportionally represent all of the various types of colleges and universities in the country, the data represent comparative rather than normative information.

Comparative data are published for two- and four-year colleges separately. The last set, published in 2000, included 48,999 classes for four-year colleges (957,152 students) and 20,481 classes for two-year colleges (343,663 students). Only class sizes of 10 or more are included. The data tables present means and percentile distributions for different class levels, class sizes, type of class (lecture, lab, etc.), full-time vs. part-time faculty, and most importantly, various subject fields.

Natural science, mathematics and statistics courses are generally rated lower than other fields. These courses may receive lower ratings because students' quantitative skills are often less well-developed than their verbal skills. This makes these courses more difficult to teach. Also, knowledge is growing more rapidly in natural science courses relative to other fields, such as the humanities. Teachers may feel pressured to cover increasing amounts of material within the time limits of the semester. This might make students feel rushed and confused about key material in those courses.

The comparative data also indicates small classes (less than 16 students) and classes featuring a discussion format tend to receive slightly higher ratings than larger lecture classes. This is likely because classes featuring discussions present a better learning environment for students (e.g., more individual attention).

Institutions that use the SIR II results for personnel decisions may want to consult the comparison data to take into account situational variables, such as subject fields and class size, which could affect teachers' ratings.

In interpreting SIR II results, it is vitally important that decisions are not based on small differences, sometimes referred to as the micrometer fallacy. As a prime example, the mean rating for all 48,999 classes is 3.97 on the overall evaluation of instruction item (#40), according to the four-year college comparison data. The standard deviation, a measure of the spread of scores, is .53. A significant and practical difference from the mean for all classes would be half a standard deviation, .26, which is also about 20 percentile points. Thus instructors with a mean of 3.71 or less are significantly lower than the national comparison mean. A mean of say, 3.85, while it may seem much lower, falls well within the "no difference" band around the comparison mean.

One exemplary institution uses SIR II results as one measure of teaching excellence. It requires teachers to achieve or exceed an overall SIR II score of 3.80 or higher, with a minimum expectation of 3.5. (This places these teachers in the approximate top two-thirds or top three-quarters, respectively, with the national comparative data.) Workshops address each SIR II scale so teachers can learn how to strengthen their teaching practice. Of course, this college includes other measures of teaching excellence in addition to student ratings.

My book, *Reflective Faculty Evaluation*, discusses several methods of measuring teaching effectiveness, as well as the assessment of other faculty functions.

ETS makes *Guidelines for the Use of Results of the Student Instructional Report II (SIR II)* available to all users. These guidelines briefly touch on some of the recommendations made in this paper, as well as additional suggestions for the SIR II report. One guideline, in particular, recommends ways in which institutions may employ standardized procedures for administering the SIR II forms in class, which is especially important when the results are to be used in personnel decisions.

This issue paper was written by John A. Centra, currently Professor Emeritus and Research Professor at Syracuse University. Mr. Centra was a former Senior Research Psychologist with ETS. He can be contacted via email at jacentra@syr.edu.

For more information call **1-800-745-0269** or visit **www.ets.org/sirII-1.html**.

References

Centra, J.A. *The Student Instructional Report: Its Development and Uses*. SIR Report No. 1. Princeton, N.J.: Educational Testing Service, 1972.

Centra, J.A. "The Effectiveness of Student Feedback in Modifying College Instruction." *Journal of Educational Psychology*, 1973, 65(3), 395-401. Also, SIR Report No. 2, 1972.

Centra, J.A. "Self-Ratings of College Teachers: A Comparison with Student Ratings." *Journal of Educational Measurement*, 1973, 10(4), 287-295. Also, SIR Report No. 2, 1972.

Centra, J.A. "The Relationship Between Student and Alumni Ratings of Teachers." *Educational and Psychological Measurement*, 1974, 34(2), 321-326. Also, SIR Report No. 3, 1973.

Centra, J.A. "Student Ratings of Instruction and Their Relationship to Student Learning." *American Educational Research Journal*, 1977, 14(1), 17-24. Also, SIR Report No. 4, 1976.

Centra, J.A. *Determining Faculty Effectiveness: Assessing Teaching, Research, and Service For Personnel Decisions and Improvement*. San Francisco: Jossey-Bass, 1979.

Centra, J.A. *Reflective Faculty Evaluation*. San Francisco: Jossey-Bass, 1993.

Centra, J.A. "The Use of the Teaching Portfolio and Student Evaluations for Summative Evaluations." *Journal of Higher Education*. 1994, 65, 555-570. Also, SIR Report No. 6, 1992.

Centra, J.A. *The Development of the Student Instructional Report II*. Princeton, N.J., Educational Testing Service, 1998.

Centra, J.A. "Will Teachers Receive Higher Student Evaluations by Giving Higher Grades and Less Coursework?" *Research in Higher Education*, Oct. 2003. Also, SIR Report No. 10, 2002.

Centra, J.A. and Gaubatz, N.B. "Is There Gender Bias in Student Evaluations of Teaching?" *Journal of Higher Education*, 2000, 70(1), 17-33. Also, SIR Report No. 8, 1999.

Centra, J.A and Gaubatz, N.B. *Student Perceptions of Learning and Instructional Effectiveness in College Courses: A Validity Study of SIR II*. SIR Report No. 9, 2002.

Comparative Data: Four-Year Colleges and Universities, Academic Years 1995-2000. Princeton, N.J.: Educational Testing Service, 2000.

Comparative Data: Community Colleges, Two-Year Colleges and Technical Institutions. Princeton, N.J.: Educational Testing Service, 2000.