ETS SIR II™

Listening. Learning. Leading.®

Differences in Responses to the Student Instructional Report: Is It Bias?

John A. Centra

May 19, 2009

Differences in Responses to the *Student Instructional Report*: Is It Bias?

John A. Centra

Suppose women teachers received higher ratings than men teachers on the *Student Instructional Report* (SIR II™). Does that mean that women are superior teachers or are students biased in their favor in their SIR II responses? Actually, the question cannot be answered without more information. If students also learned more in classes taught by women, as measured by achievement tests, for example, then the student ratings are evidence of more effective teaching by women and there is no bias in their SIR II responses. If that learning did not occur, then we could conclude that students were favoring women teachers in their ratings. As a matter of fact, women and men teachers receive generally similar ratings on SIR II, as well as on a measure of student learning. As will be shown by the study described in this report, bias as we define it did not occur.

Bias: What is it?

One commonly accepted definition of bias in student evaluations is as follows: bias exists when a student, teacher, or course characteristic influences the evaluations made, either positively or negatively, but is unrelated to any criteria of good teaching, such as increased student learning. In this report, I will discuss studies with the SIR II or SIR that shed light on possible bias in the ratings. Sources of possible bias include gender of the teachers or students, class size, expected grade, academic fields, course difficulty or workload, two-year vs. four-year colleges, teaching experience, and required courses vs. others. Where and when possible, student learning has been taken into account by including student responses to the Student Outcomes Scale (Scale F) of SIR II. That scale includes student ratings of their progress toward course objectives, increase in learning, increase in interest in the subject matter, the extent to which the course helped them to think independently about the subject, and the extent to which the course actively involved them in what they learned. The responses on the outcomes scale, therefore, provide a measure of student self-reported learning, which is being used as a proxy measure in lieu of achievement test results. Scale F thus provides a learning measure that could be used to analyze possible bias across a large number of diverse subject area courses; a common achievement test is not available across diverse courses and therefore would not be possible for the analyses. Moreover, self-reported learning has been shown to be related to achievement test scores (Pike,1995).

The results of the various studies are summarized in the accompanying table. Listed are:

- the sources of differences or possible bias
- whether significant or practical differences exist across classes
- do these differences reflect bias
- the action, if any, an institution should take

The SIR was introduced in 1972 and studies of it, several of which are reported here, were conducted over the next two decades. SIR II updated and improved the original SIR and was introduced in 1995. Some of the studies reported here with the SIR II included about 49,000 classes with 10 or more students in each class; the SIR analyses were most often based on random samples of 14,410 students and 8800 teachers.

Gender of Instructors and Students, and Interaction Between Them

Past studies of gender bias in the evaluation of college instructors often used small samples and did not fully resolve the issue, although the preponderance of evidence seems to point to little bias (Feldman, 1993). Our large sample study with SIR II indicated some gender preferences, particularly with female student ratings of female teachers. But the differences in ratings, though statistically significant, were not large and would have little practical importance (Centra and Gaubatz, 2000). The data in this study included 741 classes, each of which had an enrollment of at least 10 female and 10 male students. Eight academic fields were studied as well as the total sample. Important variables such as the method of teaching and the different areas of instruction (the SIR II scales) were also part of the analysis. The occasional higher evaluations received by female teachers from female students (and males in Natural Sciences) are likely due to differences in teaching styles. The women in this study were more likely than men teachers to use discussion rather than a lecture method and, as a group, appeared to be more nurturing to their students. However, there was no evidence that their students learned more, so their more favorable ratings may be due to their style of teaching. Therefore, as indicated in the enclosed table, we conclude from our study that the small differences due to instructor or student gender, or the interaction between them, were not evidence of bias. Therefore, no action need be taken by institutions in interpreting their SIR II results for women and men teachers.

Differences by Class Size

According to the *Comparative Data Guide for Four-Year Colleges* (2006), smaller classes generally receive higher ratings than larger classes on SIR II. Specifically, classes of 15 or fewer have a median score of 4.15 on the overall evaluation of instruction, item #40 (based on 29,589 classes). Classes of 16 to 35 have a median score of 4.05 (71,248 classes); classes of 36 to 100 have a median of 4.00 (10,120 classes); and classes of over 100 have a median of 3.89 (290 classes). The differences between the very smallest and largest is about a third of a standard deviation, which institutions may well want to take into account in interpreting SIR II results. Use of the Comparative Data Guide would enable institutions to do so. But are these differences between small and large classes evidence of bias? Not if one takes into account self-reported student learning. Students apparently say they learn more in smaller classes, as evidenced by their responses to the course outcomes scale. Classes of 15 or fewer have a median score of 3.88 on Course Outcomes, Scale F, while classes of 100 or more have a median of 3.49 on the same scale, a gap of .39, which is significant. Therefore, smaller classes received higher ratings AND students report learning more in them, so the SIR II ratings are not biased; perhaps smaller classes provide a better learning environment as well as having more upper-class majors in them.

Expected Grade in a Course

Some teachers believe that a sure way to win student approval, and thus higher ratings on the SIR II, is to give students undeservedly high grades. SIR II asks students to report the grade they expect to receive in the course; students do not know their actual final grade at the time they respond to the SIR II. Expected grade has correlated about .20 with the overall evaluation of instructional effectiveness, a correlation very similar to what Feldman (1997) reported in his review of other studies. This relationship between expected grades and ratings has other explanations other than the quid pro quo one of grading leniency causing higher student evaluations. Foremost is the validity explanation: that when students receive high grades in a course it is a reflection of how well they have learned, and thus they should rate the course and teacher highly. But is there in addition some bias as well on the part of students? The question of bias can only be answered by controlling for other variables that may affect course evaluations, especially a measure of student learning.

We studied this issue with a large diverse sample of SIR II classes (Centra, 2003) and found that the average expected grade that instructors had given in their courses had little effect on students' evaluations. A regression analysis controlled for student self-reports of learning (the Course Outcomes Scale), as well as other

variables that could affect student evaluation of courses such as teaching method and class size. The same result was found in our analyses within each of eight subject areas: Natural Science ($N = 10,590$), Social Science (9787), Humanities (12,943), Engineering and Technology (6397), Business (5446), Education (3693), Health (2465), Fine Arts (3171). Although previous studies had indicated an average correlation of .20 between expected grades and overall student evaluations, this study had a correlation of .11, most likely because the overall evaluation item on SIR II asks students to rate the contribution of the quality of instruction to their learning (most other instruments ask students to rate the teacher or instruction generally). Furthermore, our study correlated the mean course evaluations of students across classes with the mean expected grade in those classes; many other studies correlated individual student responses.

Academic Field Differences

As the *Comparative Data Guide for Four-Year Colleges* (2006) indicates, there are some differences among the various fields of study in the evaluations students give to their courses. The *Guide* provides data for 69 different fields, although some of the tables for the fields are based on fewer than 100 classes (e.g., Classics, Dental Services) and others include several thousand classes. It is important to keep in mind that the data are comparative rather than normative because it is not based on a random sample of institutions or classes. Even so, examining the data does provide an indication of how the larger fields of study, which are the only ones we will compare, are evaluated by students on the SIR II.

Courses in the natural sciences, mathematics, engineering, and computer science had an average mean of 3.87 on the overall evaluation of instruction item. This was about a third of a standard deviation less than courses in the humanities (English, history, languages) which had a mean of 4.04. Other subject fields fell generally between those two areas in their overall evaluation of courses. The 2000 comparative guide booklets, based on half as many classes and fewer subject fields, had similar results, as did analyses by Feldman (1978) and Cashin (1990) of other student evaluation instruments. A third of a standard deviation does not have much practical significance, but institutions may still want to use the comparison data for subject fields when evaluating faculty for personnel decisions. Teachers in the natural sciences (physics and chemistry, in particular), mathematics, engineering, and computer sciences may prefer to be compared with other teachers in their fields rather than all other teachers.

While there were these small differences in the overall evaluation item, there were no differences between the subject fields on the Course Organization and Planning, and the Tests and Exams scales of SIR II. There were, however, differences for three scales: Faculty Student Interaction, Course Difficulty and Workload, and Communication (Educational Testing Service, 2006, 1990). An earlier study conducted at five colleges indicated that science and math/statistics areas were rated slightly lower than other subject areas. Those results seem to explain these findings of the way that teachers and students viewed their courses: while teachers in the natural sciences thought the level of difficulty and pace of their courses were appropriate, students disagreed. Those teachers also thought that students did not put enough effort into their courses; again students disagreed (Centra, 1973).

Various reasons may be given as to why quantitative courses receive slightly lower ratings than humanities courses. Students' quantitative skills are often less advanced than their verbal skills. Natural science courses may also be more difficult to teach because knowledge is growing more rapidly in those areas and teachers feel pressured to cover increasing amounts of material; as a result, students may feel rushed and confused about important material. And finally, natural science teachers may spend more time seeking funds and doing research than humanities professors; their teaching may suffer as a result (Cashin, 1990). For these reasons, the slightly lower evaluations received by teachers of the natural sciences, mathematics/statistics, and computer sciences may not be evidence of bias, but a reflection of the effectiveness of teaching.

Two-Year vs. Four-Year Colleges

The 2006 *Comparative Data Guide for Community Colleges, Two-Year Colleges and Technical Institutions* includes results for 66,481 classes. For four-year colleges and universities, 117,132 classes are included. Both samples have at least 10 or more students in each class. In comparing these large samples, we find very small differences in the results for the items and scales in SIR II. For example, on item 40, the overall evaluation of instruction, the means are 4.04 and 3.99 for two-year colleges and four-year colleges, respectively. The scale scores are very similar, with the largest difference for Student Effort and Involvement. Two-year colleges were .15 higher on that scale, suggesting that students in those colleges say they put a little more effort into their courses than do students at four-year colleges and universities, not surprisingly, since they are probably a little less prepared for college-level work.

These results tell us that the differences in SIR II results between two- and four-year colleges are small enough to be ignored. So why publish comparative data for each group? It's because faculty members and administrators generally prefer to compare SIR II results for their college to other institutions in their own category.

Teaching Experience

The first year of teaching, as in other professions, can be expected to be important for improving the skills and knowledge learned in professional programs. This is, in fact, supported by students' evaluations on SIR. Our 1976 study of a sample of 8863 teachers in four-year colleges indicate that students rated teachers in their first year of teaching lower than teachers who had been teaching one or two years, and considerably lower than those with three or more years of teaching (Centra and Creech, 1976). The differences between teachers in their first year and the groups with additional years were all significant. A similar, though not as dramatic a finding, is displayed with the 2006 SIR II comparative data. A sample of 1539 teaching assistants, some portion of which were undoubtedly in their first year of teaching, had a mean score of 3.83 on item 40, the overall evaluation of the quality of teaching in a course. Higher ranked teachers, assistant professors and above, scored about a third of a standard deviation higher on the same overall evaluation item. Most likely, the gap would be greater if years of experience instead of rank was used to classify the teachers.

That teachers in their first year of teaching receive lower ratings is not evidence of bias but probably a reflection of their lower quality teaching, something that could be expected to improve as teachers gain additional experience. Most of these younger teachers would hopefully be open to development and improvement efforts, such as provided by the compendium of suggestions which are tied into SIR II results. The collection is titled "Enhancing Your Teaching Though Use of the SIR II Report: Suggestions for Improvement," and is available on the web for SIR II users (go to ETS.org, then to Higher Education, SIR II, and finally the compendium).

An assumption often made is that part-time teachers, because they generally lack as much teaching experience as full-time faculty members, will be much less effective as teachers. The data do not back this up. According to the comparative data guides, full-time faculty members scored 4.00 on overall evaluation of the quality of instruction, while part-time faculty members scored only slightly less at 3.96. The results for two-year colleges were similar: 4.07 for full-time faculty members and 4.02 for part-time teachers. These results are too small to be considered of practical significance.

Required Courses vs. Others

The vast majority of two- and four-year colleges have some required courses that students must take regardless of their major field of study. These courses, usually part of a general or liberal education requirement, are typically taken in a student's first and second years of study. Because students are often less motivated to take these required college courses, as compared to courses in their major, minor, or as an elective, they may be expected to rate them less favorably. This is verified by our study of SIR data, which found a slight

bias against college-required courses (Centra and Creech, 1976). The difference is not great, and is understandable given that students may be required to take course work in areas in which they have no interest or background: for example, humanities or social science majors required to take a basic mathematics or natural science course, which are often large classes as well. Given this possibility, and even though the ratings for college-required courses are only slightly lower than others, institutions may want to take the reason for the course into account in viewing SIR II results.

CONCLUSION

There is little evidence of bias in the studies and analyses we have done with SIR II and SIR results. Still, instructors could be teaching courses which have a combination of characteristics that put them at a disadvantage in the ratings they receive. For example, someone teaching a very large mathematics or natural science course that fulfills a college general education requirement may have a difficult time being rated as well as other teachers. That person is teaching a course that students could have less motivation to take, and in a large class setting, that is a less desirable teaching environment. Teachers and administrators need to be aware of situations such as these.

References

Cashin, W.E. (1990) Students do rate different academic fields differently. In M. Theall and J. Franklin (eds.), Student Ratings of Instruction: Issues for Improving Practice, *New Directions for Teaching and Learning*. 43. San Francisco: Jossey-Bass.

Centra, J.A. (1972) Effectiveness of student feedback in modifying college instruction. *Journal of Educational Psychology*. 65 (3) 305-401. Also SIR Research Report #2.

Centra, J.A. (1993) Reflective Faculty Evaluation. 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*, V. 44. No. 5.

Centra, J.A. and Creech, (1976) The Relationship Between Student, Teacher and Course Characteristics and Student Ratings of Teacher Effectiveness. PR_761. Educational Testing Service, Princeton, N.J.

Centra, J.A. and Gaubatz, N.B. (2003) Is there Gender Bias in Student Evaluations of Teaching? *Journal of Higher Education* 70 (1) 17-33.

Educational Testing Service (2000), Student Instructional Report II, Comparative Data. Academic Years 1995-2000. Four- Year Colleges and Universities.

Educational Testing Service (2000), Student Instructional Report II, Comparative Data Academic Years 1995-2000. Community Colleges, Two-Year Colleges and Technical Institutions.

Educational Testing Service (2006), Student Instructional Report II, Comparative Data, Four-Year Colleges and Universities.

Educational Testing Service (2006), Student Instructional Report II, Comparative Data, Community Colleges, Two-Year Colleges and Technical Institutions.

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

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 (2), 151-211.

Feldman, K.A. (1997) Identifying exemplary teachers and teaching: Evidence from student ratings. In Perry, R.P. and Smart, J.C. (eds.), Effective Teaching in Higher Education: Research and Practice. New York: Agathon. 368-395.

Pike, G.R. (1995) The relationship between self-reports of college experiences and test scores. *Research in Higher Education*, 36, 1-22.

Differences In Student Evaluations of Instruction: Is It Bias?

Based on Analyses of the Student Instructional Report II (SIR II) and the Student Instructional Report (SIR)[1]

John A Centra

Source of Differences/Bias	Significant/Practical Difference Across Classes?[2]	Is It Bias?	What Action Should the Institution Take?
Gender:			
of Instructor	No	No	None
of Student	No	No	None
Interaction	Rare	No	None
Class Size	Slight Advantage to Small Classes	No	May Use Comparative Data
Expected Grade (Grading Leniency)	Yes	No	None
Course Difficulty Or Workload	Yes	No	Courses rated "Just Right" were best
Academic Fields	Slight Disadvantage to Natural Sci., Math, Engineering and Computer Sci.	Maybe	May Use Comparative Data
Two-Year vs. Four-Year Colleges	No Difference	No	Separate Comparative Data is available
Teaching Experience	Teachers in their First Year Are Rated Lower	No	None. Be informed.
Required Courses vs. others	Slight Disadvantage to College Requirements Not Part of Major	Maybe	Take course requirement into account

[1] Some analyses based on 49,000 four-year college classes of 10 or more students for SIR II; for SIR, based on random samples of 14,410 students and 8800 teachers.

[2] Based on half a standard deviation or more; slight difference was about one-fourth a standard deviation.