IDEA Editorial Note #2 • Response to "Bias Against Female Instructors"

Ken Ryalls, President • Steve Benton, Senior Research Officer
Jason Barr, Researcher • Dan Li, Research Associate
The IDEA Center

A recent headline from *Inside Higher Education* proclaimed, "Bias Against Female Instructors," a hasty conclusion reached by Colleen Flaherty who reviewed a study by Anne Boring, Kellie Ottoboni, and Philip Stark (2016). The Boring et al. article was published in *ScienceOpen Research*, an online publication that undergoes minimal review. The editor takes about a week to check ethical standards, methods, and integrity in the presentation of results, and the article is published after the author pays a fee. Peer review is post-publication. To quote the video on the Web site, "You [the author] can publish your work when you think it is ready" (http://about.scienceopen.com/). The post-publication peer review process is open source and transparent.

Many of the claims made in the IHE article are flagrantly misleading. Flaherty begins, "There is mounting evidence suggesting that student evaluations of teaching are unreliable." Either the IHE author misunderstands the concept of reliability or is unaware of or ignoring the relevant literature. Reliability refers to consistency, and well-constructed student ratings of instruction (SRI) have a great deal of it (see review by Benton & Cashin, 2014, and IDEA Paper No. 50). Actually, SRI within the same class tend to be highly consistent in students' own ratings, in ratings over students within the same class, and in ratings of the same instructor across multiple courses. The article by Boring et al. questions the *validity* of student ratings, not their reliability.

Next, Flaherty concludes that, "the association between evaluations and perceived instructor gender in both the U.S. and French data sets is largely statistically significant." In point of fact, the correlations between average student evaluations of teaching (SET) scores and gender in the French sample of 1,194 students across six courses ranged from .04 to .11, *none significant at the .05 level*. So, **gender explained no more than 1% of the variance in ratings**.

Putting aside Flaherty's missteps, we have several concerns about the Boring et al. study itself:

1. What exactly did the SET measure?
2. What validity and reliability evidence is there for the learning measure?
3. What effect did researcher expectancy effects have in the U.S. study?
4. What effect did having *only* male lecturers have on French students?
5. Many of the correlations reported are very weak and non-significant.
6. Why should we assume assignment of instructors to sections in the French sample was "as if at random"?
7. Correlation is not causation.
8. How generalizable are these findings?

What Exactly did the SET Measure?

Boring et al. report the results of two studies conducted on separate samples, one from six courses offered in France, the other from one course in the U.S. The authors claim to have found gender bias in both SET instruments. So, it is logical to ask, "What exactly did those SET measure?" As an aside, we prefer the term student ratings of instruction to "student evaluations of teaching" (SET). Using the term "rating" rather than "evaluation" distinguishes those who provide data from those who interpret or evaluate it (Benton & Cashin, 2011).

Regarding the French sample, the only possible answer is we don't know what the SET measured. Readers are simply told it included closed-ended and open-ended questions. But, we know nothing of the nature of the questions. Did they ask students about how much they learned in the class, how well the instructor employed certain teaching methods, how well the instructor taught, or whether the instructor cared? We are told only that the "item that attracts the most attention… is the *overall score*, which is treated as a summary of other items" (p. 6). So, is the overall score a sum of all the closed-ended items or a single item? If it is a summative score, problems arise when items tapping into dissimilar constructs are averaged (Cashin, 1999). If it is a single item, it would certainly help to know how

it was worded. But, again, no information is provided about any of the items on the SET nor whether they correlate with any relevant measure of teaching effectiveness. **So, we really do not know what construct is being correlated with instructor gender.**

The SET used in the U.S. sample was described previously in MacNell, Driscoll, and Hunt (2014). The 15-item instrument was comprised of Likert-type items inviting students to respond from 1 = *Strongly disagree* to 5 = *Strongly agree*. Six items were intended to measure effectiveness (e.g., professionalism, knowledge, objectivity), six were for interpersonal traits (e.g., respect, enthusiasm, warmth), two were included for communication skills, and one was "to evaluate the instructor's overall quality as a teacher." No information about the exact wording of the items was provided. Moreover, the authors provided no theoretical explanation for item development or whether the "student ratings index" correlates with any other relevant measures. The only statistically significant differences ($p < .05$) were found on three traits—fairness, praise, and promptness. We contend that those three characteristics are not necessarily an indication of overall teaching effectiveness. Moreover, interpersonal traits are ones most likely to be rife with gender bias. To find gender bias when asking about personal traits is uninteresting at best. *Of greater interest is the finding that the one item that measured "overall quality" was not significantly different by gender.*

So, in the French study we do not know exactly what aspect of teaching effectiveness is being correlated with instructor gender. In the U.S. study, we know that overall teaching quality is NOT associated with instructor gender.

Where is the Validity and Reliability Evidence for the Learning Measure?

The learning measure in the French study was a final exam written by the respective course professor. We are given no information about the quality or number of items, whether they are objective or essay questions. No information is provided about their psychometric properties, how reliable they are, whether they have validity as measures within their respective domains. Yet, when scores on the final do not correlate significantly with SET, we are directed by the authors to conclude the failure rests with the SET for not being a good measure of teaching effectiveness. It seems prudent to ask whether the final exam is actually a good measure of how much students learned or of how well the instructor taught.

What Effect Did Research Expectancy Effects Have in the U.S. Study?

Researcher expectancy effects can occur when those carrying out a study know what is expected. Regarding the U.S. sample, MacNell et al. (2014) reported that, "All instructors were aware of the study and cooperated fully." So, in other words the instructors knew that in one section they were identified as a person of the opposite gender. The authors should have employed a double-blind procedure so that neither instructor would have known which section was the "perceived gender." Blind procedures are basic to good social science, and to ignore them is surprising, to say the least, and suggests a motive for the desired result from the research. As Krathwohl (1993) points out, "Researchers or their assistants may inadvertently tip the scales in favor of an experimental treatment in a variety of ways…for example with encouragement and clues" (p. 468). Notably, two of the three items that were significantly different between the perceived female and male sections were associated with encouragement (i.e., praise) and objectivity (i.e., fairness), which could have been subject to inadvertent expectancy effects, because the instructors might have responded differently on the discussion boards across their two sections. Given the complexity in the online interaction with individual students, it would have been difficult, if not impossible, for the instructors to "maintain consistency in teaching style" (p. 6). Although the authors apparently want us to assume the instructors behaved exactly the same way in each section they taught—the one for their actual gender and the one for their perceived gender—they provide no information about what actually occurred in those course sections. They could have performed a content analysis of the discussion boards, but they did not.

Related to this issue is the fact that the participating students were enrolled in an anthropology/sociology course. Was gender bias a topic in the course? Did the instructors inadvertently or expressly share views about gender bias?

What Effect Did Having Only Male Lecturers Have on French Students?

In the French study, each of six courses had one male professor who delivered the lectures to groups of approximately 900 students in total. Course sections, comprised of 10-24 students each, were taught by a mixture of male and female instructors. So, for all 1,194 students, across all courses, *the primary source of content knowledge was a male figure*. We find that peculiar and unlike what students in the U.S. might experience. Whereas women hold approximately 35% of academic positions in France, in the U.S. they hold

nearly 50%, and perhaps even a higher percentage in sociology.[1] What effect did always having a male knowledge source have on students whose section was taught by a female instructor? Did this cause students to view female faculty as less credible?

Many of the Correlations Reported Are Very Weak and Non-significant

As mentioned previously, the correlations between SET and instructor gender in the French sample were weak and non-significant. Even the correlations between SET and gender concordance ranged from -.11 to .18. The weak correlations showing French male students gave slightly higher ratings to French male instructors is in direct contrast to the Centra and Gaubatz (2000) study of U.S. students. Those authors found that female instructors received slightly higher ratings, especially from female students. Centra and Gaubatz concluded, as we do regarding the current study, that the effect of instructor gender is so small it should most likely not affect personnel decisions, *especially if student ratings are not the only measure of teaching effectiveness*. As we at IDEA have consistently pointed out, student ratings should never be the only measure of teaching effectiveness.

Why Should We Assume Assignment of Instructors to Course Sections was "as if at random"?

Boring et al. report that the enrollment process in the French sample did not allow students to select individual instructors. They argue that the assignment of instructors to course sections was, therefore, "as if at random, *forming a natural experiment*" (p. 6). But, this is different than saying all instructors were assigned to sections randomly. We can assume, then, they were not. Even if they were, it would not be the same as assigning students to sections randomly. Why make such an issue of this? Once again, it is a basic premise of social science that seems to have been ignored. If instructors were not randomly assigned (which they were not) and if students were not randomly assigned (which they were not), any differences found between male and female instructors in SET could be attributed to student characteristics that differed across course sections. Student characteristics such as motivation, work habits (Benton & Cashin, 2011, 2014), and background preparation (Benton, Li, Brown, Guo, & Sullivan, 2015) are positively correlated with student ratings of instruction. It could be, then, that the weak correlations reported between SET and gender actually reflect variance across course sections due to student characteristics.

Correlation is Not Causation

We believe it is helpful to remember what all scientists learn early in their education—correlation demonstrates association, not causal relationships. In the previous paragraph, we have mentioned one alternative explanation for the weak correlation between instructor gender and SET. But, it could also be the case that instructor gender, in this study, is related to something else (e.g., years of teaching experience) that is also related to SET. We actually know nothing about the instructors other than their sex. Given that women hold fewer of the academic positions in France than men, one might wonder whether the women in the French sample had less teaching experience. Only about 20% of women hold full professorships in Europe, whereas about 40% of assistant and associate professors are female (Ministry of Higher Education and Research of France, 2013). Teaching experience is, in fact, related to ratings (Braskamp & Ory, 1994), as it should be.

How Generalizable Are These Findings?

The samples used in Boring et al. came from single institutions, one across six courses in France, the other from a single course in the U.S. In contrast, Centra and Gaubatz (2000) analyzed student ratings data from 741 courses from eight discipline groups across 21 two - and four-year U.S. institutions. The SET they used, Student Instructional Report (SIR) II, developed by the Educational Testing Service in 1998, has high reliability and validity (Centra, 1998).

Why Do Journalists Continue to Highlight Validity Threats to SET Based on Isolated Studies Fraught with Validity Threats?

Given the onslaught of negative articles published recently about SET (e.g., Asher, 2013; Berrett, 2015; Mulhere, 2014; Wieman, 2015; Zimmerman, 2014), we pause to ask why? What motivates journalists to post headlines about studies that find weak evidence of bias in student ratings and ignore hundreds that document ample evidence of reliability and validity? The old adage, "If it bleeds, it leads," comes to mind.

---

[1] According to NCES (https://nces.ed.gov/programs/digest/d14/tables/dt14_315.10.asp), in 2013 48.8% of college faculty in the U.S. (*n* = 1,544,060) were female. According to the Ministry of Higher Education and Research of France (http://cache.media.enseignementsup-recherche.gouv.fr/file/2013/43/6/NI_MESR_13_07_3_266436.pdf), in 2011-2012, 21.4% of Full Professors (Professeurs titulaires et associés) and 42.8% of MCF titulaires et associés (Assistant/Associate Professors, https://en.wikipedia.org/wiki/Academic_ranks_in_France) were women. There are 21,116 Professeurs titulaires et associés and 37,298 MCF titulaires et associés in 2011-2012. So, (21116 * 0.214 + 37298 * 0.428)/(21116 + 37298) = 0.3506

We challenge editors to take greater care in objectively evaluating the meaningfulness of research findings, or at the very least to allow counterpoints into the discussion for fairness.

Concluding Remarks

In conclusion, the Boring et al. study, reported in IHE falls short of other studies investigating gender and student ratings. In studies of ratings of actual teachers there is only a very weak relationship that favors female instructors (Centra, 2009; Feldman, 1993). This is not to say that gender bias does not exist. We grant that it can be found in all walks of life and professions. But a single study fraught with confounding variables and weak correlations should not be cause for alarm. The gender differences in student ratings reported previously (e.g., Centra & Gaubatz, 2000; Feldman, 1992, 1993) and in Boring et al. (2016) are not large and should not greatly affect teaching evaluations *especially if SET are not the only measure of teaching effectiveness*. But, even if they are the only measure, this study shows gender contributes only about 1% of the variance in student ratings.

As has always been the case, the IDEA Center recommends that student ratings count no more than 30% to 50% of the overall teaching evaluation. Moreover, ratings and other evaluative material (e.g., student products, peer observations, course documents) should be collected from at least 6 to 8 classes before summative decisions are made about an individual faculty member.

Lastly, the Boring et al. study boldly pointed out flaws in SET without also mentioning those of other widely used measures of teaching effectiveness. Peer ratings typically represent observations of only one person on one occasion, and so they are subjected to the individual biases of the person doing the review. External recognition (e.g., rewards, invitations to speak about one's teaching) may or may not be related to good classroom instruction. Participation in professional development is commendable only if it is connected with better teaching. Embedded assessments (student products, writing samples) suffer from their own validity and reliability threats. Faculty self-evaluation is subjective—some are good at "blowing their own horn" whereas others are, perhaps for cultural reasons, more reluctant to speak highly of themselves. Finally, what constitutes "hazardous duty" (e.g., large classes, online courses) is sometimes difficult to define.

The bottom line is all measures have their shortcomings. The solution is to base evaluation of teaching effectiveness on multiple measures collected across multiple occasions and to direct criticism of evaluation where it usually belongs—on the process itself.

REFERENCES

Asher, L. (2013, October 27). When students rate teachers, standards drop. *The Wall Street Journal*. Downloaded 5/4/2015: http://www.wsj.com/news/articles/SB10001424052702304176904579115971990673400

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

Benton, S. L., & Cashin, W. E. (2014). Student ratings of instruction in college and university courses. In Michael B. Paulsen (Ed.), *Higher Education: Handbook of Theory & Research*, Vol. 29 (pp. 279-326). Dordrecht, The Netherlands: Springer.

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

Berrett, D. (2015, December 18). Scholars take aim at student evaluations' 'air of objectivity'. *The Chronicle of Higher Education*. Downloaded 5/4/2015: http://chronicle.com/article/Scholars-Take-Aim-at-Student/148859/

Boring, A., Ottoboni, K., & Stark, Ph. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*. https://www.scienceopen.com/document/vid/818d8ec0-5908-47d8-86b4-5dc38f04b23e

Cashin, W. E. (1999). Student ratings of teaching: Uses and misuses. In P. Seldin (Ed.), Changing practices in evaluating teaching (pp. 25-44). Bolton, MA: Anker Publishing Co., Inc.

Centra, J. A. (1998). Development of The Student Instructional Report II. Princeton, NJ: Educational Testing Service.

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.

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.

Krathwohl, D. R. (1993). *Methods of educational and social science research*. New York, NY: Longman.

MacNell, L., Driscoll, A., & Hunt, A. N. (2014). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, DOI 10.1007/s10755-014-9313-4.

Ministry of Higher Education and Research of France. (2013). Les personnels enseignants de l'enseignement superieur sous tutelle du MESR—2011-2012. *Note d' Information*, *7*, 1-8.

Mulhere, K. (2014, December 10). Students praise male professors. *Inside Higher Education*. Downloaded, 5/4/2015 https://www.insidehighered.com/news/2014/12/10/study-finds-gender-perception-affects-evaluations

Wieman, C. (2015). A better way to evaluate undergraduate teaching. *Change*, *Jan./Feb.*, 7-15.

Zimmerman, J. (2014). The real scandal behind the Yale course Web site. *The Washington Post*, Jan. 24. https://www.washingtonpost.com/opinions/the-real-scandal-behind-the-yale-course-web-site/2014/01/24/f719ef56-8449-11e3-9dd4-e7278db80d86_story.html

T: 800.255.2757
T: 785.320.2400

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