Can Women Teach Math (and be Promoted)? A Meta-Analysis of Gender Differences across Student Evaluations of Teaching

By: Amanda J Felkey, Lake Forest College; Cassondra Batz-Barbarich, Lake Forest College

Online Appendix

Complete Meta Data Collection Procedures

Systematic Review of Publications

To ensure a comprehensive examination of this question, we searched for articles in several databases, including EconLit, Academic Search Complete, Business Source Elite, Education Full Text, ERIC, APA PsycArticles, APA PsycInfo, using the search terms “("gender" OR "sex" OR "men" OR "women" OR "ethnic*" OR "color" OR "race") AND ("college" OR "higher ed*" OR "university") AND ("faculty” OR “instructor” OR “teacher" OR “professor”) AND (“evaluation*” OR “rating*”) NOT ("med*" OR "hospital" "disease" OR "brain damage" OR "disorders" OR "health" OR "elementary" OR "high school").” (Note that asterisks enable a search to pull texts that begin with the relevant keyword but have different endings. For instance, “rating*” means that the database will search for “rating” as well as “ratings). The “NOT” function in the search enabled us to exclude texts that were pulled in our search but were not relevant to our primary research question. Limitations were also placed on the search regarding geographic region, language, and year of publication. Only studies conducted in the United States and written in the English language were included. Further only studies completed during the last one hundred years (1920-2020) were included. The search results in 6,078 unique articles when exact duplicates were removed. Of this initial set, 829 were found to be potentially relevant and deemed worthy of further review. The decision as to whether an article was potentially relevant was based on the article title, keywords, and abstract.

To ensure comprehensiveness in our search, we conducted a secondary search using Google Scholar utilizing the search terms: ("gender" OR "sex" OR "men" OR "women" OR "ethnic*" OR "color" OR "race") AND ("college" OR "higher ed*" OR "university") AND ("faculty” OR “instructor” OR “teacher" OR “professor”) AND (“evaluation*” OR “rating*”). The search was also limited to the English language and to the years 1920 through 2020. There were 1,470,000 hits utilizing these search terms and limiters. The first 100 pages with 20 results per page were searched. We limited the search to the first 100 pages because after approximately 20 pages, very few, if any, additional articles were being deemed potentially relevant. From this search, an additional 266 unique articles were pulled as being potentially relevant based on its title, keywords, or abstract.

Together, our initial searches pulled 1,095 potentially relevant articles. For these articles, inclusion and exclusion decisions were made following an examination of the complete article. The decisions regarding inclusion and exclusion were made utilizing trained undergraduate research assistants. Every article was coded by 2 researchers who then compared their inclusion/exclusion decisions. Discrepancies were discussed with the lead author on the project and a final decision on the inclusion or exclusion of an article was made with the approval of the lead author.

Inclusion/Exclusion

Studies were included in the meta-analysis if they met five criteria. First, a study was included in the final analyses if it contained some type of student evaluation regarding an instructor at the college or university undergraduate level. Evaluations of all aspects of teaching and the course were included. However, evaluations that were not specific to an individual instructor (e.g., departmental evaluation, team-taught courses) were excluded as were peer or chair evaluations. Studies that contained evaluations for educators at the elementary or high school level were excluded as were evaluations completed samples that were comprised of a majority of graduate students (i.e., more than 50%). Second, the study was included if the evaluation was for an instructor in one of the following subject areas: (a) economics (b) anthropology/archeology, (c) geography, (d) history, (e) law, (f) linguistics, (g) politics, (h) psychology, and (i) sociology. Studies that collapsed across one or more of these areas were included if the included areas were only from this list. Third, for the study to be included in the final analyses, the study had to contain gender as a variable of interest. This was determined by the mention of gender in the title, abstract, main hypotheses, or method as a moderator that was recorded or examined in the analyses. Including studies that looked at gender in a secondary analysis or as a moderator allowed us to ensure there was not partiality to studies with significant differences. Fourth, the study had to provide enough information to compute a Cohen's *d*, which was calculated in one of two ways: (a) directly from the sample sizes, means, and standard deviations of the gender subgroups if they were provided or (b) indirectly by converting parametric statistics (e.g., correlations, *F* test, or *t* test) using standard conversion formulas (e.g., Rosenthal, 1994). If a study met the other inclusion criteria but did not meet these criteria, we contacted the authors via e-mail for the relevant information to compute Cohen's *d*. Fifth, the study was included if it did not violate the assumption of independent samples – a core assumption of meta-analyses. In other words, a study that used the same data, or a portion of the same data, as another study that had been included was excluded

From these articles, we collected an additional set articles through an ancestral and forward search of the included articles. The ancestral search required an examination of the literature review section and references of the included articles. Again, potential relevance of these articles was based on the article title or the description of the article in the text of its source. The forward search involved utilizing the Web of Science database to search for articles that had cited the included articles in their own reference sections. Again, potential relevance was based on the title of the article, keywords, and abstract.

In sum, based on the initial search utilizing the academic databases and Google Scholar as well as the ancestral and forward searches, 15 total articles were included in the analyses. These articles provided 39 effect sizes comprising a total of 83,025 unique SETs—54,280 evaluating male faculty and 28,745 for female faculty.

Coding

All included articles were coded in their entirety for analysis including several factors that may moderate the relationship between gender and STEs. The coded variables included the variables necessary to calculate an effect size including the social science discipline and the relevant statistical information (mean, standard deviation, sample size for evaluations of women and men). Additionally, included studies were coded based on study characteristics as well as sample characteristics (e.g., year data collected, percent of female students).

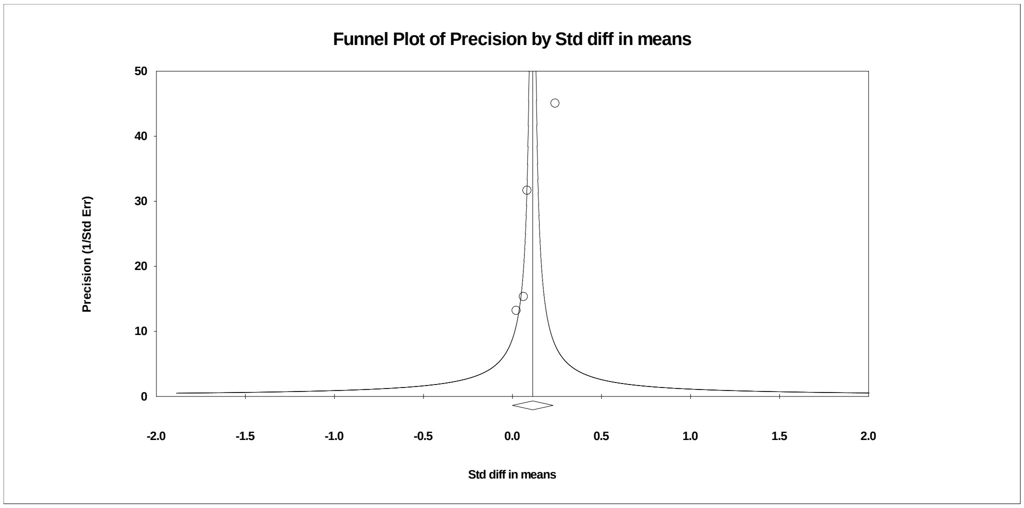
Four coders coded each of the included articles. The coders were undergraduate students who had been extensively trained on the coding manual and processes through a number of examples that were completed together. To ensure accuracy of the coding, two coders would meet for consensus meetings and then again in their assigned team of four. Differences or disagreements in the coding were mainly due to human error such as overlooking relevant information and typos.

Test for Publication Bias

We used two methods to determine whether our results suffer from publication bias. First, we examine funnel plots where the vertical axis representing the precision of the study and the horizontal axis represents the effect size. Larger studies appear at the top and smaller at the bottom. An asymmetrical distribution around the center lines indicates the potential for publication bias. If visual examination suggests bias, then we conducted the trim and fill analysis (Borenstein et al., 2009). This method imputes studies to create a symmetric funnel and determines the impact of the bias. Publication bias is the difference between the observed effect size and that calculated from the meta-analytic *d* value based on the imputed studies.

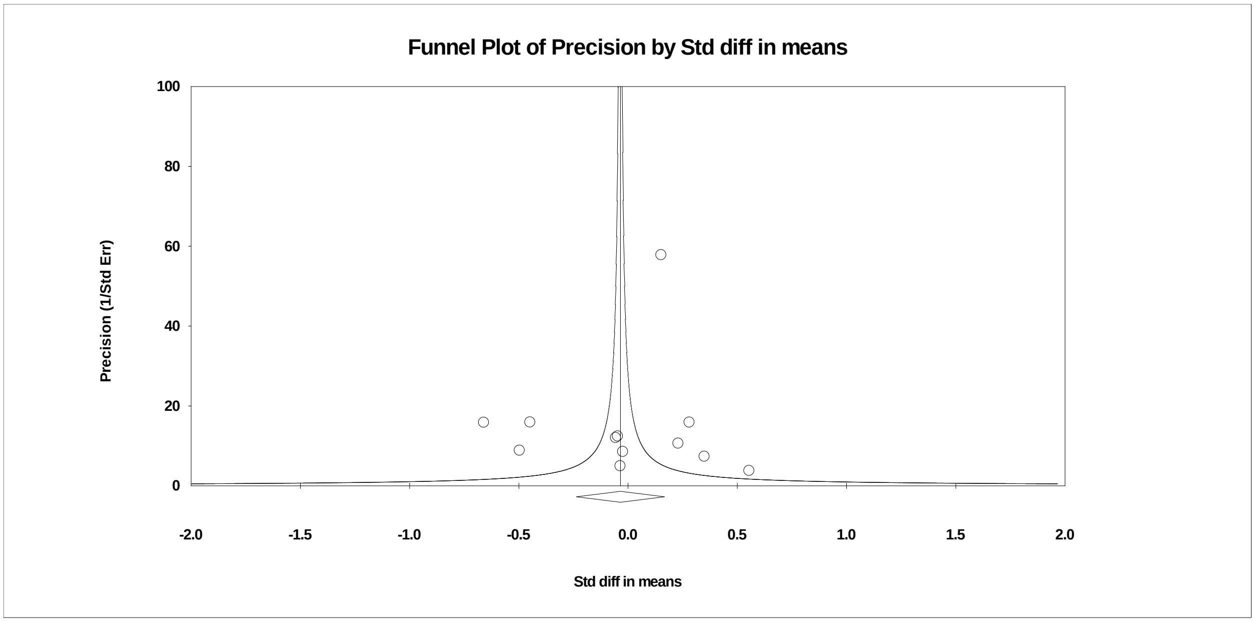
Based on funnel plot for economics (see Figure 1), there appear to be some potential for publication bias based on the visual imbalance and higher concentration of studies on the left side of the mean as compared to the right. Following up this visual examination with the Duval and Tweedie's Trim-and-Fill analysis that imputes these hypothetically missing studies to create a balanced plot, adding them to the analysis and recalculates the effect based on their addition. Based on the results of the trim-and-fill analyses, there was evidence for publication bias among economics studies, but bias that *deflated* the overall effect. The adjusted value accounting for bias reflected a *greater* difference between men and women in favor of men than what our analyses suggested (*d* = 0.137, 95% CI = [0.038, 0.236])

Figure 1. Funnel Plot for Economics Studies



Based on funnel plot for the social sciences (see Figure 2), there does not appear to be publication bias based on the visual balance of studies on both sides of the mean. Based on the results of the trim-and-fill analyses, there was no evidence for publication bias among the remaining social science studies (*d* = -0.035, 95% CI = [−0.237, 0.166]).

Figure 2. Funnel Plot for Other Social Science Studies



The results of these analysis would lead us to remain confident in the significant gender difference found for economics as well as the insignificant result for the remaining social sciences. However, publication bias analyses are not conclusive, and should be interpreted with care.

Articles Included in Meta-Analysis

Abel, M. H., & Meltzer, A. L. (2007). Student ratings of a male and female professors’ lecture on sex discrimination in the workforce. Sex Roles, 57(3-4), 173-180.

Anderson, K. J., & Smith, G. (2005). Students’ preconceptions of professors: Benefits and barriers according to ethnicity and gender. Hispanic Journal of Behavioral Sciences, 27(2), 184-201.

Arbuckle, J., & Williams, B. D. (2003). Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. Sex Roles, 49(9-10), 507-516.

Basow, S. A. (1990). Effects of teacher expressiveness: Mediated by teacher sex-typing?. Journal of Educational Psychology, 82(3), 599.

Blackwell, L. V. (2010). The Role of Gender Expectations and Organizational Citizenship Behaviors on Teaching Evaluations.

Bonitz, V. S. (2011). Student evaluation of teaching: Individual differences and bias effects.

Buser, W., Hayter, J., & Marshall, E. C. (2019, May). Gender bias and temporal effects in standard evaluations of teaching. In AEA Papers and Proceedings (Vol. 109, pp. 261-65).

Cain, K. M., Wilkowski, B. M., Barlett, C. P., Boyle, C. D., & Meier, B. P. (2018). Do we see eye to eye? Moderators of correspondence between student and faculty evaluations of day-to-day teaching. Teaching of Psychology, 45(2), 107-114.

DeFrain, E. (2016). An analysis of differences in non-instructional factors affecting teacher-course evaluations over time and across disciplines.

Joye, S., & Wilson, J. H. (2015). Professor age and gender affect student perceptions and grades. Journal of the Scholarship of Teaching and Learning, 126-138.

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. Innovative Higher Education, 40(4), 291-303.

McPherson, M. A., Jewell, R. T., & Kim, M. (2009). What determines student evaluation scores? A random effects analysis of undergraduate economics classes. Eastern economic journal, 35(1), 37-51.

Mitchell, K. M., & Martin, J. (2018). Gender bias in student evaluations. PS: Political Science & Politics, 51(3), 648-652.

Rosen, A. S. (2018). Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors. com data. Assessment & Evaluation in Higher Education, 43(1), 31-44.

Saunders, K. T., & Saunders, P. (1999). The influence of instructor gender on learning and instructor ratings. Atlantic Economic Journal, 27(4), 460-473.