Sex Differences in Higher Education Course Evaluations: A Meta-Analysis

Stephanie C. Payne, Nathanael L. Keiser, Xiaohong (Violet) Xu, & Kathleen Monks
Department of Psychology

Abstract

This study provides an update to Feldman’s (1993) meta-analysis of sex differences in student evaluations of professors. Initial results indicate minimal differences ($d = .04$), slightly in favor of women, contradicting initial speculation. Not enough articles have been coded yet to contrast STEM vs. nonSTEM as well as various other proposed moderators (e.g., level of analysis, interaction of student and instructor sex). Coding continues.

Introduction

- There are approximately 1.5 million faculty members in various higher education institutions across the U.S. (U.S. Department of Education, 2013).
- Despite a rise in the number of women faculty (48%), the percentage of women in science, technology, engineering, and mathematics (STEM) tenure-track faculty positions continues to be disproportionately lower than the percentage of men in similar positions (Nelson & Rogers, 2005).
- One explanation is that the organizational climate in the STEM disciplines is less welcoming and more “chilly” for women (Sandler, Silverberg, & Hall, 1996), with women being treated less favorably than men.
- Implicit biases about gender roles, or what people believe men and women are expected to do (Eagly & Karau, 2002) may influence student ratings.
- Course evaluations are considered biased when variables that are unrelated to good teaching (e.g., sex) are related to course evaluations (Centra, 2003; Marsh, 1987, 2007).

Method

- An initial search of PsycINFO using the search term “course evaluation” identified 1292 articles/manuscripts.
- Inclusion criteria: primary studies must contain quantitative course evaluation data for at least two college-level classes by instructors of differing sex and sufficient information to calculate relevant effect sizes.
- Although an analysis of all relevant articles from the initial search is not complete, a subset of 17 articles met the inclusion criteria and has been coded to date.
- All articles were coded by a minimum of 2 coders and good interrater agreement has been established.

Analyses

- Hunter and Schmidt’s (2004) meta-analysis methods were used to conduct a meta-analysis of $d$ values (i.e., the standardized mean difference between female and male faculty on the course evaluation).
- Corrections for two statistical artifacts were made. First, point-biserial correlations with “sex” were individually corrected using Hunter and Schmidt’s formulas (2004, p. 280) to what they would have been if each sample had a 50%-50% male-female split.
- Second, the corrected point-biserial correlations for unequal sample sizes were corrected for unreliability in the criterion using an artifact distribution.

Results

<table>
<thead>
<tr>
<th>Comparison</th>
<th>N</th>
<th>k</th>
<th>$d$</th>
<th>$SD_d$</th>
<th>$\bar{d}$</th>
<th>$SD_\bar{d}$</th>
<th>% var</th>
<th>95% CV</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female vs. Male</td>
<td>8,563</td>
<td>17</td>
<td>.04</td>
<td>.07</td>
<td>.06</td>
<td>.12</td>
<td>52%</td>
<td>-.19:.30</td>
<td>-.03:.14</td>
</tr>
</tbody>
</table>

Note. Positive $d$ values indicate that female faculty had more favorable course evaluations; $\bar{d}$ = mean sample size-weighted $d$ value; $SD_d$ = sample size-weighted observed standard deviation of $d$ values; $\bar{d}$ = mean sample size-weighted $d$ value corrected for measurement error using reliability artifact distributions; $SD_\bar{d}$ = corrected standard deviation of corrected $d$ values; % var. = percentage of variance attributable to artifacts; 95% CV = 5% and 95% credibility values, respectively; 95% CI = lower and upper bounds of the 95% confidence interval around the corrected mean $d$ value.

Conclusions and Future Research

- Hunter and Schmidt’s (2004) meta-analysis methods were used to conduct a meta-analysis of $d$ values (i.e., the standardized mean difference between female and male faculty on the course evaluation).
- Corrections for two statistical artifacts were made. First, point-biserial correlations with “sex” were individually corrected using Hunter and Schmidt’s formulas (2004, p. 280) to what they would have been if each sample had a 50%-50% male-female split.
- Second, the corrected point-biserial correlations for unequal sample sizes were corrected for unreliability in the criterion using an artifact distribution.

- Although an analysis of all relevant articles from the initial search is not complete, a subset of 17 articles met the inclusion criteria and has been coded to date.
- All articles were coded by a minimum of 2 coders and good interrater agreement has been established.

- Mean difference results for the 17 articles are shown in Table 1. Positive $d$ values indicated that female faculty had more favorable course evaluations.
- The corrected mean difference was $\bar{d} = .04$ but the confidence intervals overlapped with zero, suggesting that although female faculty had slightly more positive course evaluations, the difference may not be meaningful.

- Once additional articles are coded, proposed moderators will need to be tested.
- By better understanding the extent to which course evaluations are biased, we can determine if student-oriented implicit bias and/or search and promotion committee interventions are necessary to improve the representation of women faculty in STEM fields.