A Meta-analytic Investigation of Student Evaluations of Teaching: Re-examining Sex Differences

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

Abstract

This study reanalyzes the 23 of the 28 studies in Feldman’s (1993) meta-analysis which examined the influence of faculty sex on student evaluations of teaching (SET) and revealed no effect (p = .02; women coded higher). In this study, corrections for unequal sample sizes are made and level of analysis is examined as a moderator. Results revealed slightly stronger effect sizes favoring female professors, with the strongest effect sizes at the professor level of analysis.

Introduction

- SETs are criteria used for a number of administrative decisions including hiring, merit raises, teaching awards, promotion, and tenure (Cohen, 1980; Richardson, 2005; Seldin, 1984)
- SETs are considered biased when variables that are unrelated to good teaching (e.g., sex) are related to course evaluations (Centra, 2003; Marsh, 1987, 2007)

Feldman’s (1993) Meta-analysis

- Concerns about Feldman’s (1993) study
  - Only 28 studies published from 1936-1990
  - No corrections for unequal sample sizes
  - No corrections for criterion unreliability
  - Level of analysis not considered, violating assumptions about independent data
  - 5 studies deemed not codable

Level of Analysis

- SET data are inherently nested, with students embedded in classes, professors often times teaching more than one class
- Primary studies are conducted at one of three levels of analyses and data across levels of analysis should not be aggregated together
- Data at the student level of analysis inevitably violates assumptions of independence as students are likely to take more than one class and rate more than one professor
- Primary studies of SET data sometimes gather data across multiple semesters which also violates the assumption of independence as inevitably students and professors overlap across semesters

Analyses

- Hunter and Schmidt’s (2004) method and formulas were used to conduct a meta-analysis of r and d (i.e., the standardized mean difference between male and female faculty on the evaluation) values
- Point-biserial correlations with sex were corrected for unequal sample sizes at the professor level of analysis to what they would have been had the sample been equally split (50% male and 50% female)

Results: Reanalysis of Feldman’s (1993) meta-analysis

<table>
<thead>
<tr>
<th>Level of Analysis</th>
<th>N</th>
<th>K</th>
<th>Mean r</th>
<th>SDr</th>
<th>% var</th>
<th>95% CI</th>
<th>Mean d</th>
<th>Corrected d</th>
<th>SDr</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>25,362</td>
<td>23</td>
<td>.01</td>
<td>.06</td>
<td>14.76</td>
<td>-.03 : .04</td>
<td>.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Course</td>
<td>2933</td>
<td>3</td>
<td>.04</td>
<td>.03</td>
<td>71.65</td>
<td>-.01 : .08</td>
<td>.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professor</td>
<td>1678</td>
<td>13</td>
<td>.06</td>
<td>.14</td>
<td>38.93</td>
<td>-.02 : .13</td>
<td>.12</td>
<td>.04 : .26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professor*</td>
<td>967</td>
<td>9</td>
<td>.06</td>
<td>.10</td>
<td>35.08</td>
<td>.04 : .17</td>
<td>.14</td>
<td>.08 : .35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professor*</td>
<td>967</td>
<td>9</td>
<td>.06</td>
<td>.10</td>
<td>41.41</td>
<td>-.01 : .13</td>
<td>.14</td>
<td>.13 : .25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>20,751</td>
<td>7</td>
<td>-.002</td>
<td>.06</td>
<td>6.36</td>
<td>.06 : .05</td>
<td>-.003</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Positive r and d values indicate that female faculty had more favorable course evaluations; Mean d = mean sample size-weighted d value. Only studies with the number of male and female faculty identified; no correction applied. *Only studies with the number of male and female faculty identified, corrected to account for uneven sample sizes. Positive r values indicate that female faculty had more favorable course evaluations; r bar = mean sample size-weighted r value; SDr = sample size-weighted observed standard deviation of r values; % var = percentage of variance attributable to sampling error; 95% CI = lower and upper bounds of the 95% confidence interval around the observed correlation.

Preliminary Results for Additional Coding

<table>
<thead>
<tr>
<th>Level of Analysis</th>
<th>N</th>
<th>K</th>
<th>Mean d</th>
<th>Corrected d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>44,506</td>
<td>40</td>
<td>-.01</td>
<td>-.01</td>
</tr>
<tr>
<td>Course</td>
<td>8217</td>
<td>12</td>
<td>.10</td>
<td>.10</td>
</tr>
<tr>
<td>Professor</td>
<td>1666</td>
<td>15</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>Student</td>
<td>34,623</td>
<td>13</td>
<td>-.03</td>
<td>-.04</td>
</tr>
</tbody>
</table>

Coding challenges:
- Insufficient information to code effect size even though data obviously collected
- Authors nonresponse to requests for data
- Insufficient information to code moderators
- Multiple disciplines mixed together

Conclusions and Future Research

- To date, results are not extensively different from Feldman’s, suggesting the level of analysis and violations of data independence are not having a significant influence on effect sizes
- Virtually no difference between men and women and when there is, there appears to be a difference, students tend to rate female professors slightly higher
- Unexamined moderators remain to be tested including:
  - STEM discipline
  - Sex of the student; Interaction between sex of the student and sex of the instructor
  - Study date