scienceOPENresearch

Small samples, unreasonable generalizations, and outliers: Gender bias in student evaluation of teaching or three unhappy students?

Bob Uttl* and Victoria C. Violo

Psychology Department, Mount Royal University, 4825 Mount Royal Gate SW, Calgary, Alberta, Canada, T3E 6K6
*Corresponding author's e-mail address: uttlbob@gmail.com

Published online: 22 January 2021
Cite as: Uttl B, Violo VC. Small samples, unreasonable generalizations, and outliers: Gender bias in student evaluation of teaching or three unhappy students? *ScienceOpen Research*. 2021. (DOI: 10.14293/S2199-1006.1.SOR.2021.0001.v1)

Keywords: student evaluation of teaching, SET, small samples, outliers, generalization

ABSTRACT

In a widely cited and widely talked about study, MacNell et al. (2015) [1] examined SET ratings of one female and one male instructor, each teaching two sections of the same online course, one section under their true gender and the other section under false/opposite gender. MacNell et al. concluded that students rated perceived female instructors more harshly than perceived male instructors, demonstrating gender bias against perceived female instructors. Boring, Ottoboni, and Stark (2016) [2] re-analyzed MacNell et al.'s data and confirmed their conclusions. However, the design of MacNell et al. study is fundamentally flawed. First, MacNell et al.' section sample sizes were extremely small, ranging from 8 to 12 students. Second, MacNell et al. included only one female and one male instructor. Third, MacNell et al.'s findings depend on three outliers – three unhappy students (all in perceived female conditions) who gave their instructors the lowest possible ratings on all or nearly all SET items. We re-analyzed MacNell et al.'s data with and without the three outliers. Our analyses showed that the gender bias against perceived female instructors disappeared. Instead, students rated the actual female vs. male instructor higher, regardless of perceived gender. MacNell et al.'s study is a real-life demonstration that conclusions based on extremely small sample-sized studies are unwarranted and uninterpretable.

INTRODUCTION

In an article entitled "What's in a name: Exposing gender bias in student ratings of teaching", MacNell, Driscoll, and Hunt [1] examined whether students are biased against female faculty when completing student evaluation of teaching (SET) questionnaires. MacNell et al. examined SET ratings of one female and one male instructor teaching an online course under two conditions: when students were either truthfully told the gender of each instructor (True Gender condition) or when students were misled about their instructors' genders and told that each instructor's gender was in fact the opposite of what it was (False Gender condition). Accordingly, students evaluated a single identical female instructor under either perceived female/actual female (pF/aF) or under perceived male/actual female (pM/aF) conditions, and evaluated a single identical male instructor under either perceived female/actual male (pF/aM) or under perceived male/actual male (pM/aM) conditions. In each condition, the male and female instructors were evaluated by 8 to 12 students only. MacNell et al. stated that both instructors interacted with their students exclusively online (allowing them to mislead students about their genders) through discussion boards and emails only; graded students' work at the same time; used the same grading rubrics; and co-ordinated their grading to ensure that grading was equitable in all four sections.

MacNell et al. [1] concluded that study demonstrated gender bias in student ratings of teaching. They stated:

"Our findings show that the bias we saw here is *not* [emphasis in original] a result of gendered behavior on the part of the instructor, but of actual bias on the part of the students. Regardless of actual gender or performance, students rated the perceived female instructor significantly more harshly than the perceived male instructor, which suggests that a female instructor would have to work harder than a male to receive comparable ratings...." (p. 301)

A year later, MacNell et al.'s [1] data were re-analyzed by Boring, Ottoboni, and Stark [2] using non-parametric permutation tests rather than parametric tests used by MacNell et al. Boring et al. similarly concluded that

"The results suggests that students rate instructors more on the basis of the instructor's perceived gender than on the basis of the instructor's effectiveness. Students of the TA who is actually female did substantially better in the course, but students rated apparently male TAs higher." (p. 9)

Thus, two independents sets of three researchers analyzed MacNell et al.'s [1] data and both teams concluded that MacNell

et al.'s data were strong evidence of gender bias. However, a detailed examination of MacNell et al.'s study suggests that MacNell et al.'s conclusions are unwarranted and uninterpretable. First, MacNell et al. found no statistically significant gender difference overall (using Student Rating Index) between perceived male and perceived female (p = .128). Boring, Ottoboni, and Stark [2] confirmed the lack of statistically significant gender difference in MacNell et al.'s study using permutation test (p = .12; see their Table 8).

Second, MacNell et al.'s [1] sample of students in each of the four conditions was extremely small, ranging from only 8 to 12 students. Results based on such small samples typically have low statistical power, inflated discovery rate, inflated effect size estimation, low replicability, low generalizability, and high sensitivity to outliers [3].

Third, MacNell et al.'s [1] study included only one female and one male instructor. It is difficult to see how one could make valid generalizations about how students rate female vs. male instructors based on how students rate one particular male and one particular female instructor.

Fourth, MacNell et al.'s [1] Table 2 as well as Figure 2 suggest that the variability of SET ratings is much larger in some conditions than in other conditions, indicating the likely presence of outliers inflating variability in some but not other conditions. In fact, MacNell et al.'s data shown in Table 1, include three obvious outliers – three unhappy students who gave their instructors the lowest possible ratings on all or nearly all SET items (a familiar scenario to anyone who has ever taught such small courses). The three outliers are printed in bold in Table 1. All three occurred in perceived female conditions.

Accordingly, we examine the effect of the three outliers – three unhappy students – on MacNell et al.'s [1] findings and conclusions. Specifically, we re-analyzed MacNell et al.'s data and attempted to replicate summaries in MacNell et al.'s Table 2 and Figure 1 under two scenarios: (1) with the three outliers kept in the analyses and (2) with the three outliers removed from the data set.

METHOD

We downloaded MacNell et al.'s [1] data from http://n2t.net/ark:/b6078/d1mw2k, via the link provided in Boring, Ottoboni, and Stark [2]. We formally examined MacNell et al.'s data for outliers using Tukey's rule for identifying outliers as values more than 1.5 interquartile range from the quartiles and then re-analyzed MacNell's data with and without the three outliers plainly visible in Table 1.

Based on preliminary principal component factor analysis of their data, MacNell et al. [1] used only 12 of 15 SET items in their analyses – they excluded communicate (item 5), clear (item 14), and overall (item 15) SET items. Given the hazardous nature of conducting a principal component factor analysis on 15 variables with only 43 participants and three outliers, we used the same 12 items identified by MacNell et al. but we also examined how the mean of these 12 items correlates with the mean of all 15 items.

Specifically, we attempted to replicate MacNell et al.'s [1] summaries in Table 2 and Figure 1 and to see how these summaries would change when the three outliers were removed. Notably, neither MacNell et al. nor Boring et al. [2] mentioned outliers in their analyses of MaNell et al.'s data.

RESULTS

Figure 1, Panel A, shows the boxplot of SET ratings – the mean average of 12 items used by MacNell et al. [1]. The boxplot shows the three outliers – three students giving their instructors the lowest possible ratings on all or nearly all items. Similarly, Panel B shows the same data but for the mean average of all 15 items. The same three outliers are identified in this boxplot. Panel C shows the near identity relationship between the average of 12 items and the average of 15 items, with the correlation r = .998. This suggests that MacNell et al. would have obtained nearly identical results if they used all SET items rather than select only 12. Panel D shows the stripchart of the 12-item means for each of the four experimental conditions: pF/aF, pM/aM, pM/aF, and pF/aM. The stripchart shows that the three outliers occurred in the two perceived female conditions (i.e, pF/aF and pF/aM) and highlights the extremely small number of students in each of the four conditions, with ns ranging from 8 to 12 students.

Table 2 shows the mean student ratings for each of the 12 SET items used by MacNell et al. [1]. The top third shows the means, standard deviations, and other statistics for 12 SET items comparing the male instructor with the female instructor and comparing the perceived male and perceived female instructors as reported by MacNell et al. in their Table 2. MacNell et al. did not report actual p-values but only whether any given p-value was < .10 and < .05.

Table 2, the middle third, shows our re-analysis of MacNell et al.'s [1] data with outliers not removed. Accordingly, the values in the middle third ought to be identical to those reported by MacNell et al. and shown in the top third of the table. The values are indeed identical – we consider differences in the last significant digit as identical – to those in MacNell et al. with two notable exceptions: the values in the r^2 column comparing the male instructor with the female instructor match except for the last value in the column, and the values in the r^2 column for the perceived male and perceived female instructors do not match except the last value in the column which matches. However, the statistically significant difference between male and female, using p < .05 standard, occurred only for the perceived instructor conditions and only for fair, praise, and prompt SET items, replicating MacNell et al.'s inferential statistics conclusions.

Table 2, the bottom third, shows the identical analyses with the three outliers removed. As expected, the values change considerably except in the perceived male conditions as these did not include any outliers. First, in the actual gender conditions, the female instructor was rated higher than the male instructor on all 12 items, with the female instructor rated 0.08 to 0.54 points higher than the male instructor. For two items only, these differences were statistically significant at p < .05. Second, in the

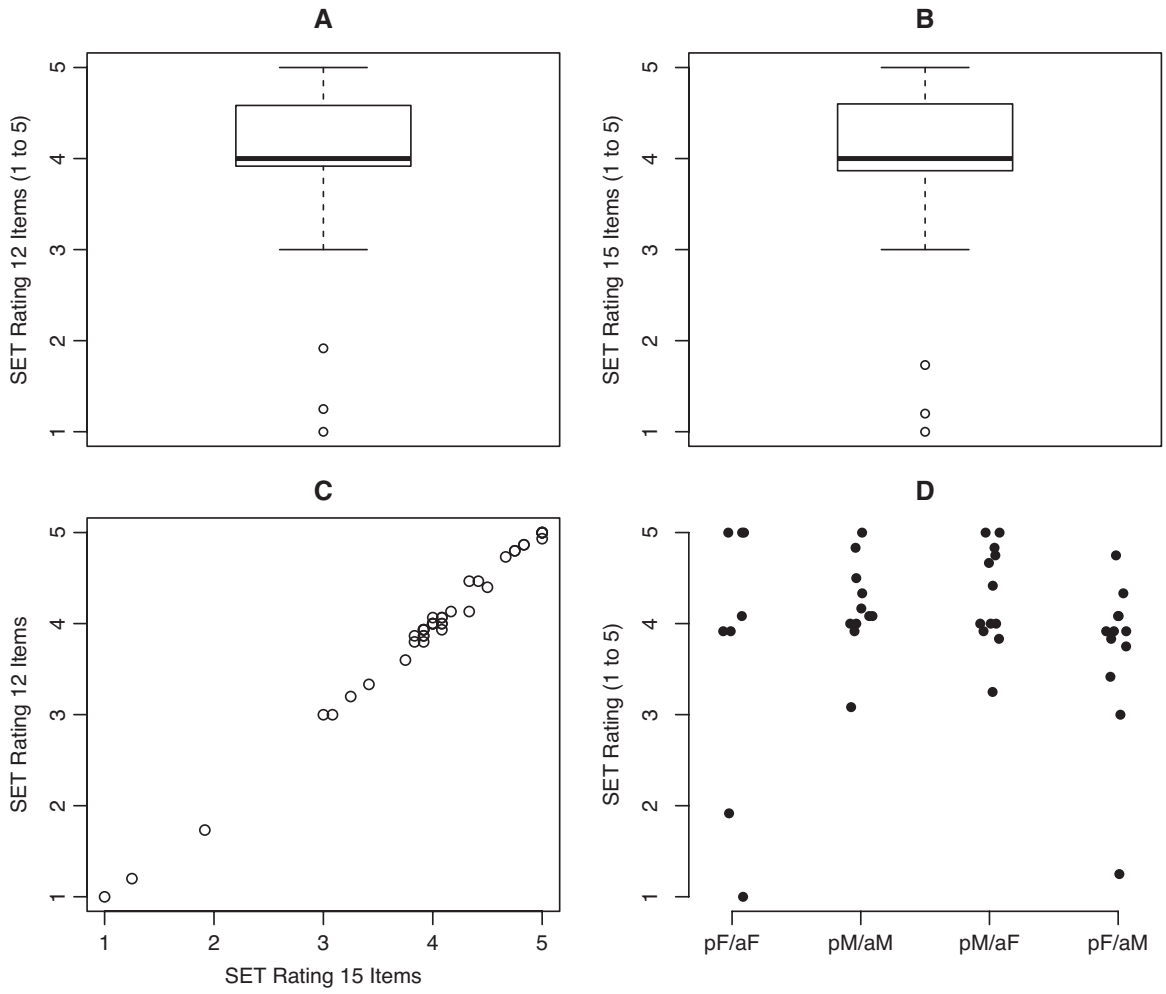

Figure 1. MacNell et al.'s [1] data. Panel A shows the boxplot of SET ratings – the mean average of 12 items used by MacNell et al. The boxplot highlights the presence of three outliers – three students giving their instructors the lowest possible rating on all or nearly all SET items. Panel B shows the same data but for the mean average for all 15 items. The same three outliers are visible. Panel C shows the near identity relationship between the average of 12 items and the average of 15 items (r = .998). Panel D shows the strip chart of the 12-item means for each of the four experimental conditions and highlights extremely small number of students in each condition. It also shows that the three outliers occurred in the two perceived female conditions.

perceived gender conditions, the female and male instructors were rated comparably, with no difference statistically significant at $p < .05$ level. Accordingly, these item level analyses showed that when the three outliers were removed, the SET effects favouring males vs. females reported by MacNell et al. [1] were wiped out and some SET effects favouring females vs. males emerged instead.

Figure 2 shows the mean SET ratings for the 12 items. Panel A shows the SET ratings for the actual male vs. female instructor and for the perceived male vs. female instructor for all data. The Actual Gender bars show the data for the actual male and actual female instructor with data collapsed across True and False Gender conditions. The Perceived Gender bars show the data for the perceived male and the perceived female instructor with the data collapsed across actual gender. This figure highlights that students rated the actual female instructor numerically higher than the actual male instructor. In contrast, when the data were collapsed across Actual Gender conditions, the students rated the perceived male instructor higher than the perceived female instructor. The Panel A directly replicates MacNell et al.'s [1] analyses reported in their Figure 2.

Figure 2, Panel B, shows SET ratings by the four experimental conditions (i.e., with no collapsing across conditions). This figure highlights that in the True Gender conditions, the male instructor was rated higher than the female instructor. In the False Gender conditions, the students rated the same female instructor who was presented as male higher than the same male instructor who was presented as female. Thus, this data pattern supports MacNell et al.'s [1] claim that it is the perception of the instructor as male vs. female that matters rather than what male vs. female instructors actually did.

However, when the three outliers are removed, the findings change. Panel C shows the identical analyses to those in Panel A but with the three outliers removed. The Actual Gender

Table 1: MacNell et al. [1] data.

	SET Item																	
Group	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**sex**	**ag**	**pg**
pM/aF	5	5	4	4	4	3	4	4	4	4	4	4	3	5	4	2	0	1
pM/aF	4	4	4	4	5	5	5	5	3	4	5	5	5	5	4	1	0	1
pM/aF	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	2	0	1
pM/aF	5	5	5	5	5	3	5	5	5	5	3	5	5	5	5	2	0	1
pM/aF	5	5	5	5	5	5	5	3	4	5	5	5	5	5	5	2	0	1
pM/aF	4	4	4	4	4	4	3	4	3	3	5	5	3	3	4	1	0	1
pM/aF	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	1	0	1
pM/aF	5	5	5	4	5	4	5	5	5	5	5	5	5	5	5	1	0	1
pM/aF	4	4	3	4	4	4	5	4	4	4	3	5	4	4	4	2	0	1
pM/aF	4	4	3	3	3	3	3	4	2	4	3	3	3	3	3	1	0	1
pM/aF	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	1	0	1
pM/aF	5	5	4	4	3	3	4	4	3	4	3	4	4	4	4	1	0	1
pM/aM	5	5	5	5	5	5	5	5	5	5	3	5	5	5	5	2	1	1
pM/aM	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	2	1	1
pM/aM	5	5	4	4	4	3	4	4	3	4	3	5	5	2	4	2	1	1
pM/aM	5	5	5	4	4	5	4	4	4	5	3	4	4	2	4	2	1	1
pM/aM	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	1	1	1
pM/aM	5	4	3	4	5	2	2	5	5	4	3	5	5	2	3	1	1	1
pM/aM	5	5	5	4	4	5	4	4	4	4	4	5	5	4	4	1	1	1
pM/aM	4	5	4	4	3	4	4	4	4	4	4	4	4	4	4	1	1	1
pM/aM	4	4	2	3	3	3	3	4	2	3	3	3	3	2	3	1	1	1
pM/aM	5	5	4	4	4	3	4	5	4	4	3	5	4	4	4	2	1	1
pM/aM	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	2	1	1
pF/aF	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
pF/aF	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**4**	**3**	**4**	**4**	**1**	**1**	**2**	**0**	**0**
pF/aF	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	1	0	0
pF/aF	5	5	4	4	4	3	4	3	3	3	4	4	5	4	4	1	0	0
pF/aF	5	5	3	4	3	2	4	5	4	3	4	4	4	4	4	2	0	0
pF/aF	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	2	0	0
pF/aF	4	4	5	5	4	3	4	5	5	3	3	4	4	3	4	1	0	0
pF/aF	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	2	0	0
pF/aM	4	4	4	4	4	4	4	4	4	4	3	4	4	4	4	2	1	0
pF/aM	5	5	5	5	5	5	5	4	4	4	5	5	5	5	5	1	1	0
pF/aM	5	5	3	4	5	4	5	4	5	5	3	4	5	5	5	2	1	0
pF/aM	5	4	4	4	4	4	4	4	4	4	4	4	4	4	4	1	1	0
pF/aM	4	4	4	4	4	4	5	4	4	4	4	4	4	4	4	2	1	0
pF/aM	4	4	4	4	4	3	4	4	4	4	4	4	4	4	4	2	1	0
pF/aM	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	1	1	0
pF/aM	5	5	4	4	4	4	4	3	3	3	4	4	4	4	4	2	1	0
pF/aM	**1**	**1**	**2**	**1**	**1**	**3**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**0**
pF/aM	4	5	4	3	4	3	4	4	3	3	4	4	5	4	4	1	1	0
pF/aM	5	5	4	3	4	4	2	2	2	2	4	4	4	1	4	2	1	0
pF/aM	4	4	3	3	3	4	4	4	4	4	4	4	3	3	3	2	1	0

Note. Group: pM/aF = perceived male/actual female, pM/aM = perceived male/actual male, pF/aF = perceived female/actual female, pF/aM = perceived female/actual male; sex: 1 = male student, 2 = female student; ag = actual gender: 0 = female, 1 = male; pg = perceived gender: 0 = female, 1 = male; SET Item: 1 = professional, 2 = respect, 3 = caring, 4 = enthusiastic, 5 = communicate, 6 = helpful, 7 = feedback, 8 = prompt, 9 = consistent, 10 = fair, 11 = responsive, 12 = praised, 13 = knowledgeable, 14 = clear, 15 = overall.

condition shows that the female instructor is rated higher than the male instructor whereas the Perceived Gender condition shows that the differences between perceived female and male instructors all but disappeared. Panel D shows the identical analyses to those in Panel B but with the three outliers removed. The data show that female instructor was rated higher than the male instructor in both the True Gender and False Gender conditions.

CONCLUSIONS

MacNell et al. [1] claimed that their findings demonstrated that students were actually biased against female vs. male instructors rather than merely being in favor of female gendered behavior. Boring, Ottoboni, and Stark [2] re-analyzed MacNell et al.'s data, confirmed MacNell et al.'s findings, and concluded that students (1) rated instructors on the basis of gender rather than teaching effectiveness, and (2) rated male teachers better

Table 2: Mean student ratings of teaching for each of the 12 items used by MacNell et al. [1]. The top third shows the data copied from MacNell et al.'s Table 2; the middle third shows our replication of MacNell et al'.s analyses; and the bottom third shows our replication of MacNell et al.'s analyses with the three outliers removed.

	Actual Gender							Perceived Gender						
SET	aF	aF	aM	aM				pF	pF	pM	pM			
Item	*M*	*SD*	*M*	*SD*	diff.	r^2	*p*	*M*	*SD*	*M*	*SD*	diff.	r^2	*p*
MacNell et al.'s analyses copied from their Table 2														
Caring	4.00	1.257	3.87	0.868	0.13	.004		3.65	1.226	4.17	0.834	-0.52	.071	
Consistent	3.80	1.322	3.70	1.020	0.10	.002		3.50	1.357	3.96	0.928	-0.47	.045	
Enthusiastic	4.05	1.191	3.78	0.850	0.27	.019		3.60	1.314	4.17	0.576	-0.57	.112	†
Fair	4.05	1.050	3.78	0.951	0.27	.018		3.50	1.192	4.26	0.619	-0.76	.188	*
Feedback	4.10	1.252	3.83	1.029	0.27	.015		3.70	1.380	4.17	0.834	-0.47	.054	
Helpful	3.65	1.309	3.83	0.834	-0.18	.008		3.50	1.192	3.96	0.928	-0.46	.049	
Knowledgeable	4.20	1.056	4.09	0.949	0.11	.003		3.95	1.191	4.30	0.765	-0.35	.038	
Praise	4.35	0.988	4.09	0.900	0.26	.020		3.85	1.089	4.52	0.665	-0.67	.153	*
Professional	4.30	1.218	4.35	0.935	-0.05	.000		4.00	1.414	4.61	0.499	-0.61	.124	†
Prompt	4.10	1.252	3.87	0.919	0.23	.013		3.55	1.356	4.35	0.573	-0.80	.191	*
Respectful	4.30	1.218	4.35	0.935	-0.05	.001		4.00	1.414	4.61	0.499	-0.61	.124	†
Responsive	4.00	1.124	3.57	0.843	0.43	.052		3.65	1.137	3.87	0.869	-0.22	.013	
Replication of MacNell et al.'s analyses														
Caring	4.00	1.257	3.87	0.869	0.13	.004	.699	3.65	1.226	4.17	0.834	-0.52	.063	.116
Consistent	3.80	1.322	3.70	1.020	0.10	.002	.776	3.50	1.357	3.96	0.928	-0.46	.040	.214
Enthusiastic	4.05	1.191	3.78	0.850	0.27	.018	.409	3.60	1.314	4.17	0.576	-0.57	.081	.083
Fair	4.05	1.050	3.78	0.951	0.27	.018	.390	3.50	1.192	4.26	0.619	-0.76	.149	.016
Feedback	4.10	1.252	3.83	1.029	0.27	.015	.442	3.70	1.380	4.17	0.834	-0.47	.045	.191
Helpful	3.65	1.309	3.83	0.834	-0.18	.007	.609	3.50	1.192	3.96	0.928	-0.46	.046	.174
Knowledgeable	4.20	1.056	4.09	0.949	0.11	.003	.716	3.95	1.191	4.30	0.765	-0.35	.033	.262
Praise	4.35	0.988	4.09	0.900	0.26	.020	.370	3.85	1.089	4.52	0.665	-0.67	.130	.023
Professional	4.30	1.218	4.35	0.935	-0.05	.001	.887	4.00	1.414	4.61	0.499	-0.61	.084	.080
Prompt	4.10	1.252	3.87	0.920	0.23	.012	.502	3.55	1.356	4.35	0.573	-0.80	.139	.022
Respectful	4.30	1.218	4.35	0.935	-0.05	.001	.887	4.00	1.414	4.61	0.499	-0.61	.084	.080
Responsive	4.00	1.124	3.57	0.843	0.43	.049	.165	3.65	1.137	3.87	0.869	-0.22	.012	.486
Re-analysis of MacNell et al.'s analyses without outliers														
Caring	4.33	0.767	3.95	0.785	0.38	.058	.133	4.06	0.748	4.17	0.834	-0.11	.005	.650
Consistent	4.11	0.963	3.82	0.853	0.29	.027	.321	3.94	0.899	3.96	0.928	-0.02	.000	.958
Enthusiastic	4.39	0.608	3.91	0.610	0.48	.139	.018	4.06	0.748	4.17	0.576	-0.11	.008	.601
Fair	4.22	0.808	3.91	0.750	0.31	.041	.216	3.76	0.903	4.26	0.619	-0.50	.101	.062
Feedback	4.44	0.705	3.95	0.844	0.49	.092	.053	4.18	0.809	4.17	0.834	0.01	.000	.992
Helpful	3.94	0.998	3.86	0.834	0.08	.002	.786	3.82	0.883	3.96	0.928	-0.14	.005	.648
Knowledgeable	4.39	0.778	4.23	0.685	0.16	.013	.495	4.29	0.686	4.30	0.765	-0.01	.000	.965
Praise	4.56	0.616	4.23	0.612	0.33	.069	.101	4.18	0.529	4.52	0.665	-0.34	.076	.075
Professional	4.67	0.485	4.50	0.598	0.17	.023	.336	4.53	0.624	4.61	0.499	-0.08	.005	.669
Prompt	4.44	0.705	4.00	0.690	0.44	.096	.053	4.00	0.866	4.35	0.573	-0.35	.058	.162
Respectful	4.67	0.485	4.50	0.598	0.17	.023	.336	4.53	0.624	4.61	0.499	-0.08	.005	.669
Responsive	4.22	0.878	3.68	0.646	0.54	.117	.038	4.00	0.707	3.87	0.869	0.13	.007	.604
N	23		20					20		23				

Note. † p < .10; * p < .05; pM/aF = perceived male/actual female, pM/aM = perceived male/actual male, pF/aF = perceived female/actual female, pF/aM = perceived female/actual male.

than female teachers even though they learned more from female teachers. However, in reality, neither MacNell et al. nor Boring et al. found the gender difference in overall SET in MacNell et al.'s data statistically significant ($p = .128$ and $p = .12$, respectively).

Our re-analyses of MacNell et al.'s [1] small-sized study demonstrates that MacNell et al.'s data do not support either MacNell et al.'s or Boring et al.'s [2] conclusions. When three outliers – three unhappy students – are removed from the data set, the data change drastically and do not support MacNell et al.'s conclusions. If the results of such small sample-sized studies of one female and one male instructor were interpretable and generalizable to all female and male instructors – and we argue that they are not, with or without outliers, and regardless of what they show – MacNell et al.'s data actually suggest that students rate male instructors lower than female instructors regardless of what they are told about their genders.

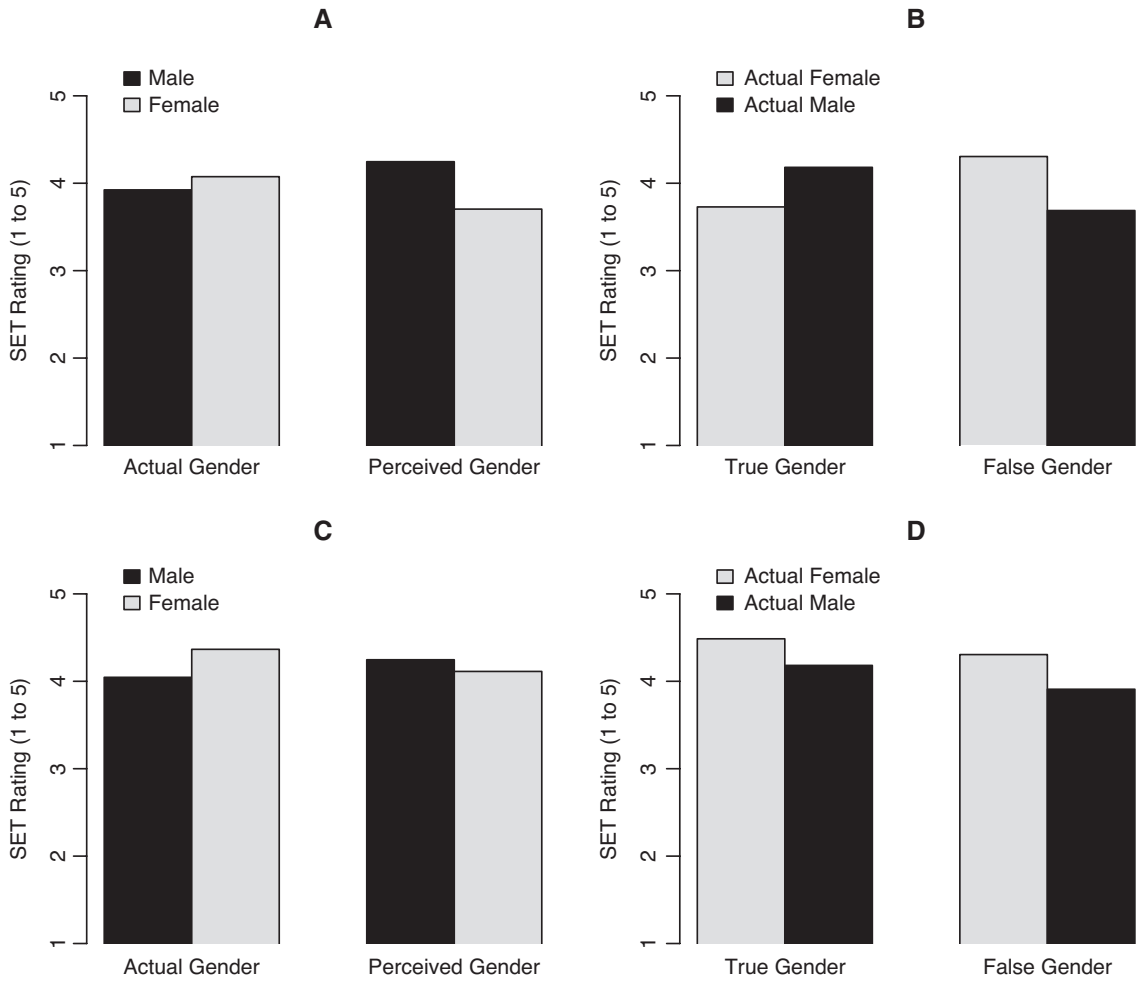

Figure 2. SET ratings for 12-item averages. Panel A shows the SET ratings for the actual male vs. female instructor and for the perceived male vs. female instructor for all data. Panel B shows the SET ratings by the four experimental conditions for all data. The instructor perceived as male received higher ratings that the instructor perceived as female. Panel C shows the SET ratings for the actual male vs. female instructor and for the perceived vs. female instructor when the three outliers are removed. Panel D shows the SET ratings by the four experimental conditions when the three outliers are removed. The actual female instructor received higher ratings than the actual male instructor, regardless of their perceived gender.

Importantly, MacNell et al.'s [1] published data highlight nothing short of the absurd practice of interpreting the mean SET ratings from a small number of students as having anything to do with the instructor. The same identical instructor (actual female) who received 4.31 SET rating in one section (pM/aF) received widely discrepant ratings of 3.73 or 4.49 in the other section (pF/aF) depending on whether or not two outliers – two unhappy students – were retained or excluded from the means, respectively. They highlight that professors ought to focus principally on students' satisfaction and ought not to do anything to lower it, for example, ought not to call students on academic dishonesty, adhere to academic standards, etc. Moreover, given the Kruger-Dunning effect [4] and SET destroying effect of one or two outliers in small classes, professors must focus on satisfying principally the least able students who would perceive the greatest discrepancy between the grades reflecting their achievement and the grades they believe their work deserves, if their grades were not inflated [5].

MacNell et al.'s [1] findings and conclusions received widespread news and social media coverage and hundreds of citations. As of March 3, 2020, MacNell et al.'s Altmetric score was 697, indicating that the article was in the 99th percentile – the top 1% of all research tracked by Altmetric. MacNell et al. has been cited 153 times within the Web of Science and 408 times within Google Scholar. We examined all of the 153 Web of Science citations to determine if the citing researchers noted MacNell et al.'s small sample sizes, unreasonable generalizations from one male and one female instructor and/or outliers. No citing article noted outliers. No citing article noted unreasonable generalization. And only one article noted small sample sizes. All citations cited MacNell et al. for evidence of gender bias against female instructors. Similarly, the Boring, Ottoboni and Stark's [2] re-analysis of MacNell et al. received

widespread attention with an Altmetric score of 525 and 243 citations on Google Scholar. We searched Google Scholar for "boring ottoboni stark outlier macnell" using full text search in an attempt to identify any article indexed by Google Scholar noting outlier effects in the MacNell et al. study. Google Scholar returned 18 results and none of them mentioned outliers in the MacNell et al.'s study.

MacNell et al.'s [1] findings of no statistically significant gender differences in overall SET ratings were recently replicated in similarly fatally flawed study by Khazan, Borden, Johnson, and Greenhaw [6]. Khazan et al. examined SET ratings of a single female TA who taught two sections of the same online course, one section under her true gender (perceived female TA) and one section under false/opposite gender (perceived male TA). Just as MacNell et al. did, Khazan et al. found no gender differences in overall SET ratings of perceived female vs. male TA (p = .73) but claimed that they found gender bias against perceived female TA nevertheless [7]. Moreover, Khazan et al. suffers from nearly identical set of fatal flaws that render their study uninterpretable and conclusions unwarranted including small samples, low statistical power, outliers, confounds, and use of a single female exemplar design [7].

MacNell et al.'s [1] study is a real-life demonstration that conclusions based on small sample-size studies are unwarranted and uninterpretable. MacNell et al.'s study design, including extremely small samples, and use of only a single woman and a single man to represent female and male professors, is simply insufficient to answer their research question. Combined with small samples, failure to examine the data, and to recognize that the summaries of the data depend critically on three outliers, three unhappy students, was only the last fatal flaw rendering the study 100% uninterpretable, and its conclusions unwarranted. In the meantime, however, the world, or at least hundreds of researchers citing MacNell et al. and Boring, Ottoboni and Stark [2], falsely believes that MacNell et al.'s study demonstrated that students are biased against female professors. It is not true; MacNell et al. did not demonstrate students' bias against female professors. If anything, their results suggest that students rate female professors higher than male professors, but it would be foolish to make that claim based on the fundamentally flawed small sample design.

ACKNOWLEDGEMENTS

We thank Amy Siegenthaler for careful reading and comments on the manuscript.

REFERENCES

[1] MacNell L, Driscoll A, Hunt A. What's in a name: exposing gender bias in student ratings of Teaching. *Innovative Higher Education*. 2015. *40*(4), 291–303. DOI: 10.1007/s10755-014-9313-4.

[2] Boring A, Ottoboni K, Stark P. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research.* 2016. 1–11. DOI: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1.

[3] Ioannidis JPA. Why most published research findings are false. *PLOS Medicine*. 2005. *2*(8), e124. DOI: 10.1371/journal.pmed.0020124.

[4] Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*. 1999. *77*(6), 1121–1134. DOI: 10.1037/0022-3514.77.6.1121.

[5] Uttl B, White CA, Gonzalez DW. Meta-analysis of faculty's teaching effectiveness: student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*. 2017. *54*, 22–42. DOI: 10.1016/j.stueduc.2016.08.007.

[6] Khazan E, Borden J, Johnson S, Greenhaw L. Examining gender bias in student evaluation of teaching for graduate teaching assistants. *NACTA Journal*. 2020. *64*, 430–435.

[7] Uttl B, Violo V. Gender bias in student evaluation of teaching or a mirage? *ScienceOpen Preprints* 2020. 1–20. DOI: 10.14293/S2199-1006.1.SOR-.PPFXXC8.v1.

COMPETING INTERESTS

Authors declare no conflicting interest.

PUBLISHING NOTES