scienceOPENresearch

Gender bias in student evaluation of teaching or a mirage?

Bob Uttl* and Victoria Violo

Mount Royal University, Alberta, Canada
*Corresponding author's e-mail address: uttlbob@gmail.com

Published online: 16 December 2021
Cite as: Uttl B, Violo V. Gender bias in student evaluation of teaching or a mirage? *ScienceOpen Research*. 2021. (DOI: 10.14293/S2199-1006.1.SOR.2021.0003.v1)

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

ABSTRACT

In a recent small sample study, Khazan et al. [1] examined SET ratings received by one female teaching (TA) assistant who assisted with teaching two sections of the same online course, one section under her true gender and one section under false/opposite gender. Khazan et al. concluded that their study demonstrated gender bias against female TA even though they found no statistical difference in SET ratings between male vs. female TA ($p = 0.73$). To claim gender bias, Khazan et al. ignored their overall findings and focused on distribution of six "negative" SET ratings and claimed, without reporting any statistical test results, that (a) female students gave more positive ratings to male TA than female TA, (b) female TA received five times as many negative ratings than the male TA, and (c) female students gave "most low" scores to female TA. We conducted the missing statistical tests and found no evidence supporting Khazan et al.'s claims. We also requested Khazan et al.'s data to formally examine them for outliers and to re-analyze the data with and without the outliers. Khazan et al. refused. We read off the data from their Figure 1 and filled in several values using the brute force, exhaustive search constrained by the summary statistics reported by Khazan et al. Our re-analysis revealed six outliers and no evidence of gender bias. In fact, when the six outliers were removed, the female TA was rated higher than male TA but non-significantly so.

"If you torture the data long enough, it will confess to anything." (Ronald Coase)

In an article titled "Examining Gender Bias in Student Evaluations of Teaching for Graduate Teaching Assistants", Khazan et al. [1] examined whether students are biased against female teaching assistants when completing student evaluations of teaching (SET) questionnaires. Students in an online upper-level undergraduate course that was taught asynchronously by a male Associate Professor were assigned to one of two teaching assistants (TAs), a male TA (TAM condition) and a female TA (TAF condition). However, unknown to students, Ms. Khazan, a female, performed TA duties of both perceived male and female TA. At the end of the course, students were asked to rate their TAs on 14-item student evaluation of teaching (SET) form using a 5-point Likert scale where 1 = Strongly disagree and 5 = Strongly agree. For purposes of their analyses, Khazan et al. converted 1 to 5 scale to −2 to +2 scale and summed the ratings across all 14 items for each student to obtained so called "cumulative evaluation score" ranging from −28 to +28. Out of 136 students invited to complete SET, 115 completed them: 60 in TAM condition and 55 in TAF condition. For some of the analyses, Khazan et al. divided students in each of the two main conditions into two subgroups depending on student gender.

Khazan et al. concluded that their study demonstrated gender bias in SET against female teaching assistants. In the Abstract, they wrote (p. 430):

Overall evaluations [SET] were positive, however, evaluations of putative male and female TAs demonstrated inconsistency across student gender. Female students demonstrated the greatest variation; 100% of females assigned to the "male" TA rated the TA positively, whereas 88% of female students assigned to the "female" TA gave positive ratings. This study corroborates literature demonstrating bias against women in SET.

Similarly, in the Summary section, Khazan et al. repeated their claim that they found gender bias against women in their study; they wrote (p. 434):

… student ratings of the putative male and female TA were uneven, with the female TA receiving five times as many negative evaluations as the male TA. The largest discrepancies in SET scores were noted between male and female student

evaluations of the putative female TA, with most low scores given to the female TA by female students…

On November 2, 2020, www.insidehighered.com featured the study on its website. In an article titled "Gender Bias in TA Evals" Colleen Flaherty [2] wrote:

"Simple in design and sobering in its results, the study found that students in an online course who had the same TA gave that TA five times as many negative evaluations when they believed that she was a woman, as compared to when they thought she was a man."

In the next paragraph, Flaherty continued and implied that Khazan et al.'s study replicated MacNell et al. [3] findings of gender bias against women; Flaherty wrote: "Female students tended to give the putative female TA the worst scores of all, paralleling the U.S.-based findings of a major 2016 study on gender bias in teacher ratings."

On November 2, 2020, Ms. Khazan announced the study and the insidehighered.com article about it in her tweet @EmilyKhazan:

It can be hard being #womaninscience, and teaching evaluations often don't help. We show gender bias in teaching reviews of graduate students…

Responses to the tweet were quick, diverse and inclusive. While some praised Khazan et al.'s work, for example, as "amazing and important work", "very cool study design", "The most bravest design", others criticized the study and pointed out that Khazan et al. "actually showed the opposite", "Tiny sample. No statistically significant main effect…", "No significant difference or in plain English no evidence of any difference or as its sometimes called bogus science," and yet others provided Ms. Khazan with career advice which we decided not to quote in this article.

On November 3, 2020, the University of Florida News featured the study in an article titled "Study reveals gender bias in TA evaluations" [4]. The first paragraph informed a reader: "At the end of the semester, the students scored the male TA higher on course evaluations, while the female TA got five times as many negative reviews."

Thus, if one were to read the abstract or the summary, www.insidehighered.com, University of Florida News, and/or Ms. Khazan's tweet, one would think that Khazan et al. article found gender bias against women, that women received 5 times as many negative evaluations as men, and that female students were principally responsible for this injustice to female TAs.

In reality, they did not. As some of the tweeters pointed out, a detailed examination of Khazan et al.'s article shows no evidence of gender bias, and thus, no evidence of five times as many negative ratings, and thus, no evidence that female students were responsible for the bias. In fact, between the Abstract and the Summary, Khazan et al. themselves wrote the following (p. 433):

The mean score for TAM [perceived male TA] (17.9, SD = 7.89) was higher than that for TAF [perceived female TA] (17.3, SD = 11.1), but the means were not statistically different ($\chi^2 = 0.12$, $p = 0.73$).

Thus, the main findings of the study was that there was no evidence of gender bias whatsoever, $p = 0.73$. Unfortunately, this main finding was not featured in their abstract nor the summary. Clearly, $p = 0.73$ is not evidence of gender differences, nor gender bias. Moreover, the difference between the two conditions is mere 0.6 points relative to pooled SD of approximately 9.5 or very tiny effects size of approximately 0.06 pooled SD. Notably, this 0.6 points difference on –28 to +28 sum across 14 items scale corresponds to tiny 0.043 (i.e., 0.6/14 = 0.043) difference on 1 to 5 Strongly disagree to Strongly agree Likert SET scale students actually rated their TAs on. Moreover, Khazan et al. did not disclose any statistical test results to support their claim that distribution of the six negative ratings out of 115 ratings overall somehow evidenced gender bias. Finally, the claim that female students were responsible for gender bias against the female TA was based on $p = 0.15$ (p. 433).

The diminutive overall difference between female and male TA ratings aside, Khazan et al. study suffers from numerous other problems that render its conclusions unwarranted and uninterpretable. First, Khazan et al. sample of students in each condition was extremely small, ranging from only 19 to 36 students. As we [5] pointed out when we re-analyzed and discussed similar prior study by MacNell et al. [3], results based on such small samples suffer from numerous problems including low statistical power, inflated discovery rate, inflated effect size estimates, low replicability, low generalizability, and high sensitivity to outliers [6].

Second, similarly to MacNell et al. [3], Khazan et al. study included only one female instructor, one exemplar of all female instructors. Accordingly, it is impossible to make any valid generalization about how students rate female vs. male instructors based on one female instructor's interactions with students.

Third, Khazan et al.'s Figure 1 show that the variability of SET ratings in some conditions is increased by presence of outliers – students who rated their TA so low as to be far removed from the bulk of the distribution. Khazan et al. did not mention the presence of these outliers, did not do any statistical test to determine whether there were any outliers in their data, did not consider possible reasons for the outliers, and did not redo their analyses with outliers removed to see how their findings would change if they did so. In fact, as noted above, Khazan et al. based their entire claim of gender bias on the distribution of the six lowest values – likely outliers – across the four conditions but reported no test to support their claim that the distribution of these outliers is in fact dependent on perceived gender of TAs.

Fourth, Khazan et al. wrote that "In addition to the TA's names, each TA had a photograph and short biography available to students…" Khazan et al. did not provide photos nor the

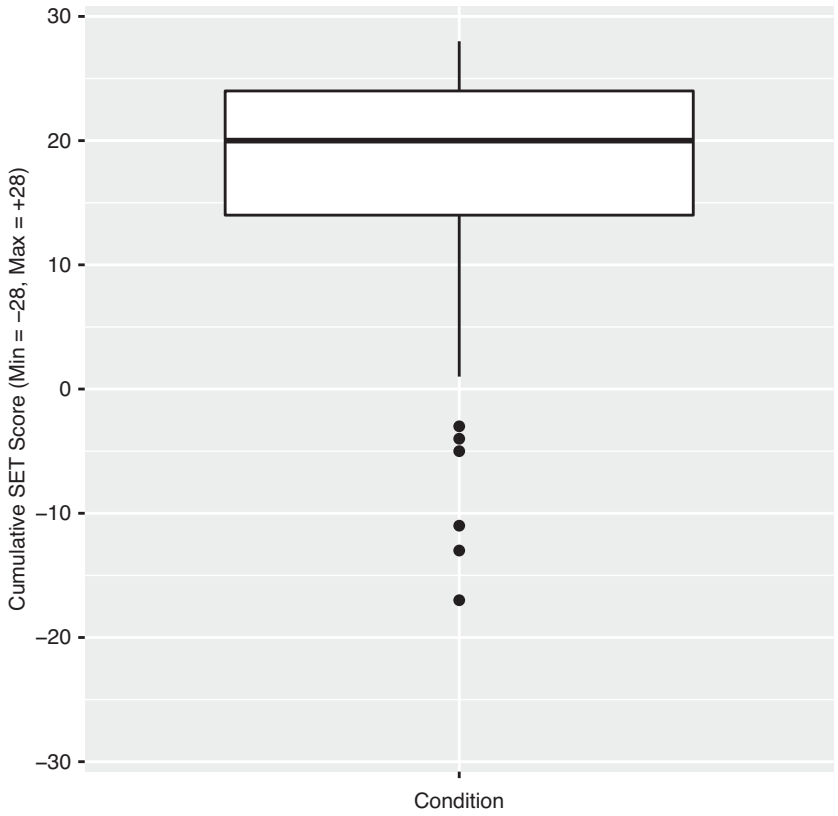

Figure 1. The boxplot of Khazan et al.'s recovered cumulative SET scores. The boxplot identifies six values as outliers, that is, the values more than 1.5 inter-quartile range from the quartiles. The outliers are the same six "negative" values used by Khazan et al. to claim gender bias against the female TA.

biographies within their article but any differences between male vs. female TA photos or biographies could be due to their looks, their perceived approachability (one of the items on SET form used by Khazan et al.), etc. In fact, the two photos are shown in Clark's [4] article describing study findings on University of Florida News (https://news.ufl.edu/2020/11/ta-bias/). It show a photo of Jesse Borden – one of the co-authors and putative male TA – looking straight into the camera and Emily Khazan examining what appears to be insects in white netting and ignoring the camera. The differences in these two photos themselves may influence students' ratings of the male vs. female TAs.

Accordingly, we examined Khazan et al. research in detail. First, we conducted the missing tests on what appears to be six outliers to see if their distribution was dependent or independent of TA genders and student genders, that is, whether there was any evidence that the female TA received more negative evaluations than male TA, and whether female students were responsible for the higher number of negative evaluations received by female TA. We realize that this amounts to what is known as data dredging, data fishing, and p-hacking given clearly non-significant overall test but we wanted to determine if there was any support whatsoever for Khazan et al.'s claims, including the claim that "the female TA [were] receiving five times as many negative evaluations as the male TA", even if data-dredged/p-hacked.

Second, we examined whether the six negative evaluations are statistical outliers as Khazan et al.'s Figure 1 suggests, and if so, whether the removal of the six outliers would change Khazan et al.'s findings and resulting conclusions. To do these analyzes, we needed access to Khazan et al.'s data. Unfortunately, Khazan et al. declined to share their data. First, they claimed that their Institutional Review Board (IRB) protocol bars them from disclosing individual respondents data. (Khazan et al., personnal communication, November 10, 2020). Second, when we pointed out that the individual respondents' data were already disclosed in their Figure 1, Dr. Greenhaw apologized for the earlier miscommunication, acknowledged that the IRB does not prevent them from sharing the data, but wrote that they – Khazan et al. – decided not to share the data nevertheless (L. Greenhaw, personnal communication, November 10, 2020). Accordingly, we read off the bulk of the data from Khazan et al. Figure 1 and filled in the remaining data points using brute force search for values that would fit the constrains on the data set by the Figure 1 and descriptive statistics provided by Khazan et al. We then emailed the recovered data set to Khazan et al. (Khazan et al., personnal communication, November 11, 2020) and asked them to confirm its accuracy or

point out which data points we recovered were incorrect and correct them. As of today, November 26, 2020, Khazan et al. did not respond. Nevertheless, we are confident that the data recovered from Khazan et al.'s Figure 1 and by the brute force search are identical or closely matching Khazan's et al's actual data that Khazan et al. refused to disclose.

METHOD

As noted above, we requested Khazan et al.'s data shown in Figure 1 from the authors. The authors initially denied access to the data because they erroneously believed that their Institutional Review Board (IRB) protocol barred them from releasing the individual level data published in their Figure 1. When this erroneous belief was cleared up, the authors nevertheless denied access to the data stating: "our research team has agreed that we prefer to not share more specific data at this time" (Dr. Greenhaw, personal communication, November 10, 2020). In response, we narrowed the request to only the data underlying the Figure 1 (Khazan et al., personal communication, November 10, 2010). We pointed out that the individual data were already visible in Figure 1, can be read off Figure 1 except a few data points overlapping with other data points, and that the authors' refusal to provide data for verification and re-analyses will merely make it more difficult to obtain the data, force us to spend more time, and to apply brute force, exhaustive algorithms to search for missing data that fit the Figure 1 and other statistics reported by the authors. Khazan et al. still chose not to share the data.

Accordingly, we proceeded to read off the data from Figure 1 using PlotDigitizer and then applied exhaustive search algorithm to determine data points that could not be read from Figure 1. Given that Khazan et al.'s reporting of their statistics contains some errors, the recovered data may be slightly different from the actual data due to these minor errors in the author's reporting of descriptive statistics. For example, the authors wrote that the mean in TAF/fs condition was 15.6 (n = 36) and that the mean in TAF/ms condition was 20.2 (n = 19) and that the overall mean in TAF condition was 17.3 (n = 55). However, this is impossible assuming normal laws of rounding numbers, and thus, the means reported in Khazan et al. are not exact and, thus, it is impossible to match them all precisely. In turn, this introduces some small degree of imprecision and uncertainty to the brute force exhaustive search. As we noted above, we attempted to get Khazan et al. to confirm the accuracy of the recovered data but they did not reply (Khazan et al., personal communication, November 11, 2020).

Khazan et al.'s recovered data are shown in Table 1. Table 1 shows the data recovered from Figure 1, for each of the four conditions: TA female/female student (TAF/fs), TA female/male student (TAF/ms), TA male/female student (TAM/fs), and TA male/male student (TAM/ms). For each data point, Table 1 indicates whether the data point was directly read off from Figure 1 or whether it was determined by the algorithm search. Table 1 shows the various summaries of the data reported by Khazan et al. The data in Table 1 match the summaries reported by Khazan et al. for each of the four groups perfectly.

Table 1: Khazan et al.'s [1] recovered data by condition and descriptive statistics, including descriptive statistics reported by Khazan et al.

Score number	TAF/female student	TAF/male student	TAM/female student	TAM/male student
1	**28**	**28**	**28**	**28**
2	**28**	**28**	**28**	**28**
3	**27**	**28**	**28**	**28**
4	**27**	**27**	**26**	**27**
5	**27**	**26**	**26**	**26**
6	**25**	**26**	**25**	**25**
7	**24**	**24**	**25**	**24**
8	**24**	**22**	**24**	**24**
9	**24**	**19**	**24**	**23**
10	**23**	**18**	**23**	**22**
11	**23**	**17**	**23**	**22**
12	**23**	**16**	**22**	**21**
13	**22**	**16**	**21**	**18**
14	**22**	**15**	**20**	**18**
15	**22**	**15**	**20**	**17**
16	**21**	**6**	**19**	**14**
17	**21**	**–3**	**19**	**11**
18	**20**	28*	**18**	**7**
19	**18**	28*	**17**	**7**
20	**17**		**15**	**4**
21	**17**		**15**	**2**
22	**17**		**14**	**–4**
23	**14**		**13**	2
24	**14**		**12**	18*
25	**14**		**12**	23*
26	**7**		**11**	
27	**6**		**10**	
28	**2**		**10**	
29	**1**		**8**	
30	**–5**		**6**	
31	**–11**		**3**	
32	**–13**		15	
33	**–17**		17	
34	25*		23*	
35	7*		24*	
36	18*			
M	15.6 [15.6]	20.2 [20.2]	18.4 [18.4]	17.4 [17.4]
SD	12.10 [12.1]	8.44	6.70 [6.69]	9.46
n	36	19	35	25

Note. The values in bold were read off from Khazan et al. Figure 1 with 100% confidence. The values in plain text were read off with high degree of certainty but not with 100% confidence. The values with the star next to them were filled using the brute force exhaustive search algorithm. The summary values in [ ] brackets are values reported by Khazan et al.

RESULTS

Statistical tests the authors did not report

Khazan et al. claimed that "Particularly noteworthy is that TAF received five times as many negative evaluations as TAM…" (p. 434). Khazan et al. reported no statistical test results demonstrating that this "five times" claim was inconsistent

with the null effect of TA gender. Using the data from Khazan et al.'s Table 1, Fisher's exact test resulted in p = 0.102, indicating no support for the claim that negative ratings (1 in TAM condition, 5 in TAF condition) were associated with TA gender. Khazan et al. also wrote that "100% of females assigned to the 'male' TA rated TA positively, whereas 88% of female students assigned to the 'female' TA gave positive rating" (p. 430). Again, Khazan et al. did not report the results of any test demonstrating that this distribution of ratings was inconsistent with null effect of TA gender. Using the data from Khazan et al. Table 1, Fisher's exact test resulted in p = 0.115, indicating no support for the claim that female students rated perceived female TA worse than perceived male TA.

Finally, Khazan et al. claimed that "the largest discrepancies in SET scores were noted between male and female student evaluations of putative female TA, with most low scores given to the female TA by female students." Again, Khazan et al. did not report the results of any tests demonstrating that there was in fact statistically significant difference in how many low scores – presumably the negative ratings – male and female students gave to perceived female (i.e., in TAF condition). Fisher's exact test resulted in p = 0.649, indicating no association between negative vs. positive ratings and student gender in TAF condition.

Re-analyses of Khazan et al.'s recovered data

Figure 1 shows the boxplot of Khazan et al.'s recovered cumulative SET scores. As expected, the boxplot identifies six values as outliers, that is, the values more than 1.5 inter-quartile range from the quartiles. The six outliers are exactly the same six values that Khazan et al. identified as "negative" evaluations, that is, the evaluations that were primarily negative, below neutral ratings.

Figure 2 shows a violin plot of cumulative SET scores. The shape of violins matches closely the violins in Khazan et al.'s Figure 1. For each condition, the violins also show individual data points, the mean of all data points, and the boxplot identifying outliers, that is, the values more than 1.5 inter-quartile range from the quartiles (edges of the box). As can be plainly seen, the six "negative"/below the midpoint of the rating scale values are all outliers. In fact, within conditions boxplots identify the same six outliers as the overall boxplot. The two means in the male TA conditions (TAM/ms, TAM/fs) are similar whereas the two means in the female TA conditions are dissimilar – TAF/fs mean is lower than TAF/ms mean, due to a number of outliers in the TAF/fs conditions. In contrast, there are much smaller differences among the medians that are not influenced by outliers.

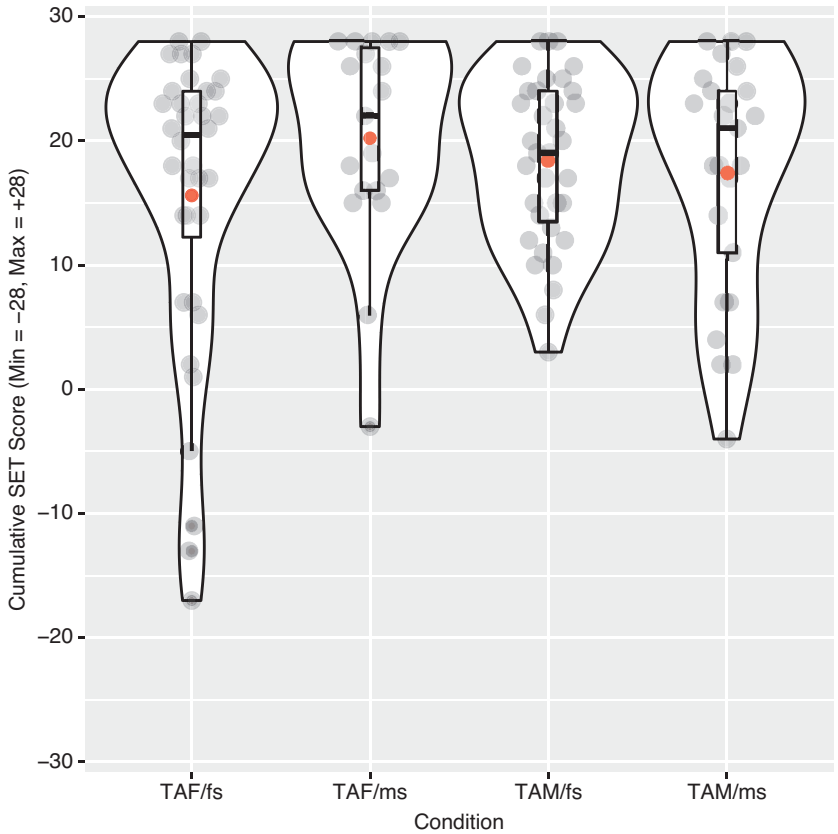

Figure 2. The violin plots of Khazan et al.'s recovered cumulative SET scores. The shape of violins matches closely the violins in Khazan et al.'s Figure 1. For each condition, violins also show individual data points, the mean of all data points (red dot), and the boxplot identifying outliers, that is, the values more than 1.5 inter-quartile range from the quartiles (edges of the box).

The overall mean was 17.6 (*SD* = 9.54). The mean scores for male (18.6, *SD* = 9.04) vs. female (17.0, *SD* = 9.85) students did not differ significantly, Kruskal-Wallis $\chi^2 = 1.162$, $p = 0.28$. The mean scores for TAM (18.0, *SD* = 7.91) vs. TAF (17.2, *SD* = 11.1) also did not differ significantly, Kruskal-Wallis $\chi^2 = 0.10$, $p = 0.75$. These results closely match those reported by Khazan et al.

Figure 3 shows the boxplot of Khazan et al.'s cumulative SET scores with the six outliers removed. The boxplots indicates no further overall outliers.

Figure 4 shows a violin plot of cumulative SET scores with the six outliers removed. The means in the two female TA conditions (TAF/fs, TAF/ms) are now higher than the means in the two male TA conditions (TAM/fs, TAM/ms), indicating that, if anything, female TAs get higher rating than male TAs. Similarly, the medians are overall higher for female TAs than for male TAs.

With the six outliers removed, the overall mean was 19.1 (*SD* = 7.32); the mean increased by 1.5 point and *SD* was reduced. The mean scores given by male (19.7, *SD* = 7.79) vs. female (18.7, *SD* = 7. 05) students was 1.0 points higher but the scores did not differ significantly, Kruskal-Wallis $\chi^2 = 1.073$, $p = 0.30$. Similarly, the mean scores for TAM (18.4, *SD* = 7.42) vs. TAF (19.9, *SD* = 7.18) conditions also did not differ significantly, Kruskal-Wallis $\chi^2 = 1.210$, $p = 0.27$, although female TA was rated 1.5 points higher than male TA, flipping the results reported by Khazan et al.

Khazan et al. chose to transpose 1 to 5 Likert scale to –2 to +2 scale and to use sums across 14 SET items to analyze their data. Above, we followed their analyzes. However, to allow a reader to appreciate magnitude of the differences and variability of the SET scores on the original SET scale, we divided Khazan et al.'s cumulative scores (sums) by 14 and added 3 to transform the data back to the original 1 to 5 scale. Thus, Figures 5 and 6 shows a violin plots for Khazan's recovered data with and without the six outliers, respectively, but on the original SET scale ranging from 1 to 5 for female and male TA.

Figure 5 (data with the outliers included) shows that female TA's mean ratings (*M* = 4.23, *SD* = 0.79) were slightly lower than male TA mean ratings (*M* = 4.28, *SD* = 0.56). In contrast, female TA's median rating was higher than male TA's median rating. Figure 6 (data with the six outliers excluded) shows that the female TA's mean rating (*M* = 4.42, *SD* = 0.51) was higher than male TA mean rating (*M* = 4.31, *SD* = 0.53).

DISCUSSION

Khazan et al.'s data reveal no evidence of gender differences: no evidence that female TA was rated differently than male TA ($p = 0.73$), no evidence that the number of negative SET evaluations was associated with TAs gender ($p = 0.102$), no evidence

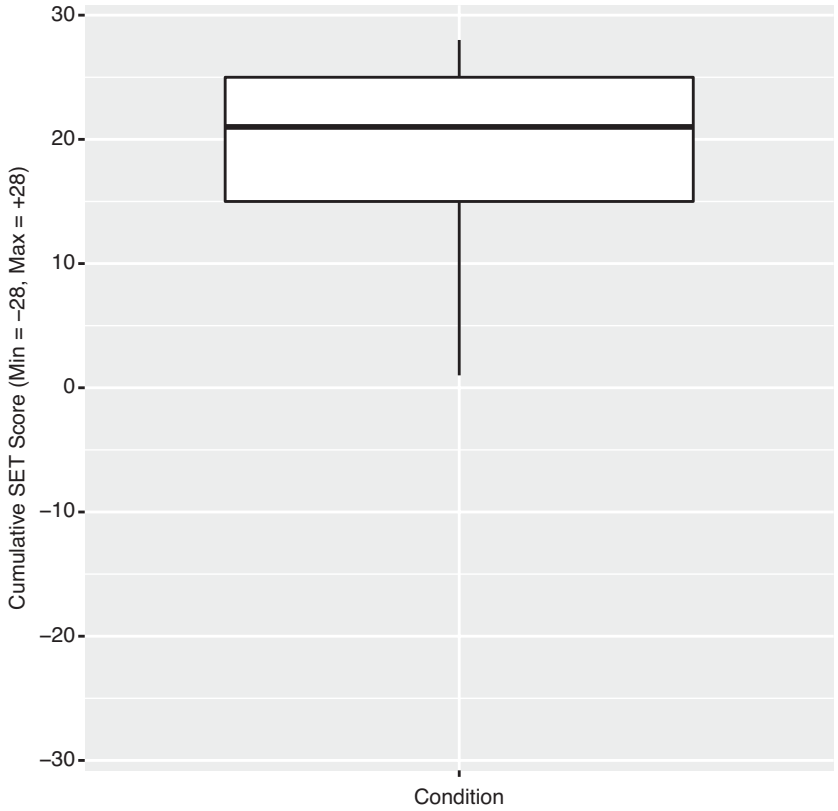

Figure 3. The boxplot of Khazan et al.'s recovered cumulative SET scores with the six outliers removed.

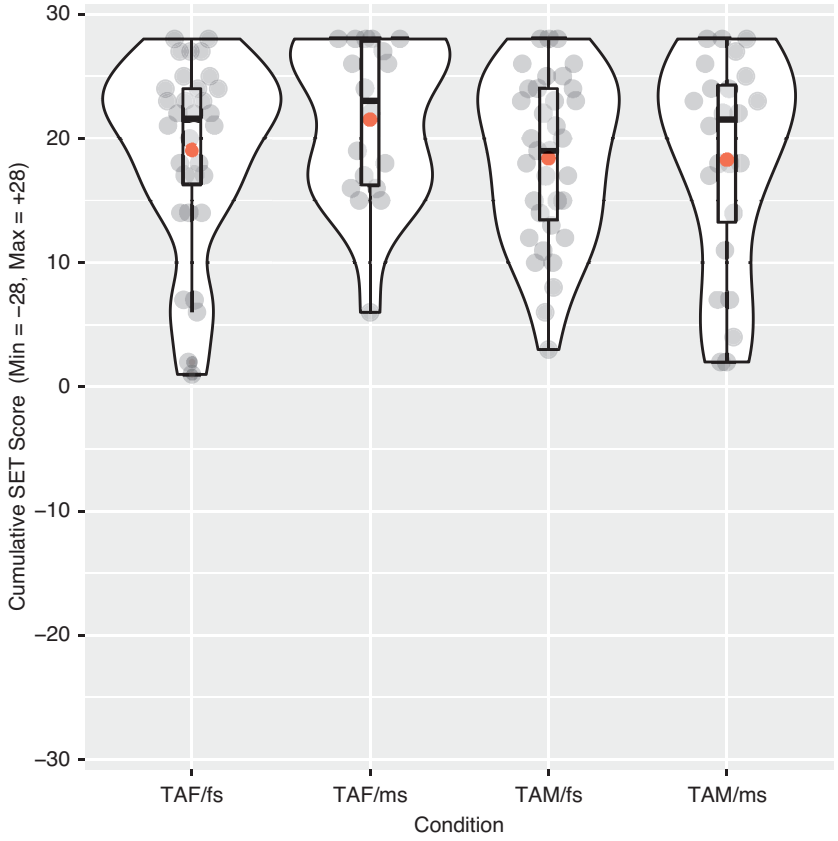

Figure 4. The violin plots of cumulative SET scores with the six outliers removed. The means in the two female TA conditions (TAF/fs, TAF/ms) are now higher than the means in the two male TA conditions (TAM/fs, TAM/ms), indicating that, if anything, female TAs get higher rating than male TAs. Similarly, the medians are also overall higher for the female TA than for the male TA.

that female students gave more negative scores to female TA than to male TA ($p = 0.115$, and no evidence that male vs. female students gave different number of negative scores to female TA ($p = 0.649$).

Khazan et al. Figure 1 plainly reveals six outliers which happen to be the "negative" evaluations. Given Khazan et al. refusal to disclose their data for our re-analyses, we were forced to recover them from Khazan et al.'s Figure 1 by reading them off and by using the brute force exhaustive search for values that we were not able to read off the Figure 1. The analysis of Khazan et al.'s recovered data (a) confirmed the six outliers, (b) revealed that female TA received slightly lower (not statistically significant) mean ratings than male TA with the outliers included, and (c) revealed that female TA received slightly higher (not statistically significant) mean ratings than male TA with the outliers excluded. When looking at median ratings, the female TA received higher ratings than male TA both with and without the six outliers included. Thus, just like with MacNell et al. [3] data, the removal of six outliers flipped the gender difference favouring male TA to gender difference favouring female TA. In contrast, the median ratings favoured female TA with and without outliers. If the gender difference was actually due to gender bias – and there is zero evidence that it is – whether or not a female TA would be advantaged or disadvantaged depends on (a) use of the means rather than medians, and (b) inclusion or exclusion of outliers. For example, in the universities that are aware of outliers and use medians or interpolated medians rather than means to summarize SET ratings, female TA would be advantaged rather than disadvantaged, although we repeat: the differences were tiny and none of them were statistically significant.

However, apart from the six outliers, Khazan et al. study suffers from several fatal flaws that render Khazan et al. data uninterpretable and any claims of gender bias unwarranted. First, Khazan et al.'s sample size was tiny. The statistical power to find gender effect of 0.2 *SD* – the slightly higher effect than that reported by Boring et al. [7] – was plainly insufficient, only about .18. If Khazan et al. wanted to have a power of at least .80 to find the effect of 0.2 *SD*, if it existed, they needed to have approximately 400 students rating each TA, that is, about seven time as many students in both male and female TA conditions as they had. Second, Khazan et al. confounded TA's gender in their study with the photos and bios of the male and female TAs. Although we do not know anything about the two bioses, the photos of the male vs. female TA were posted on the University of Florida News. The male photo showed an

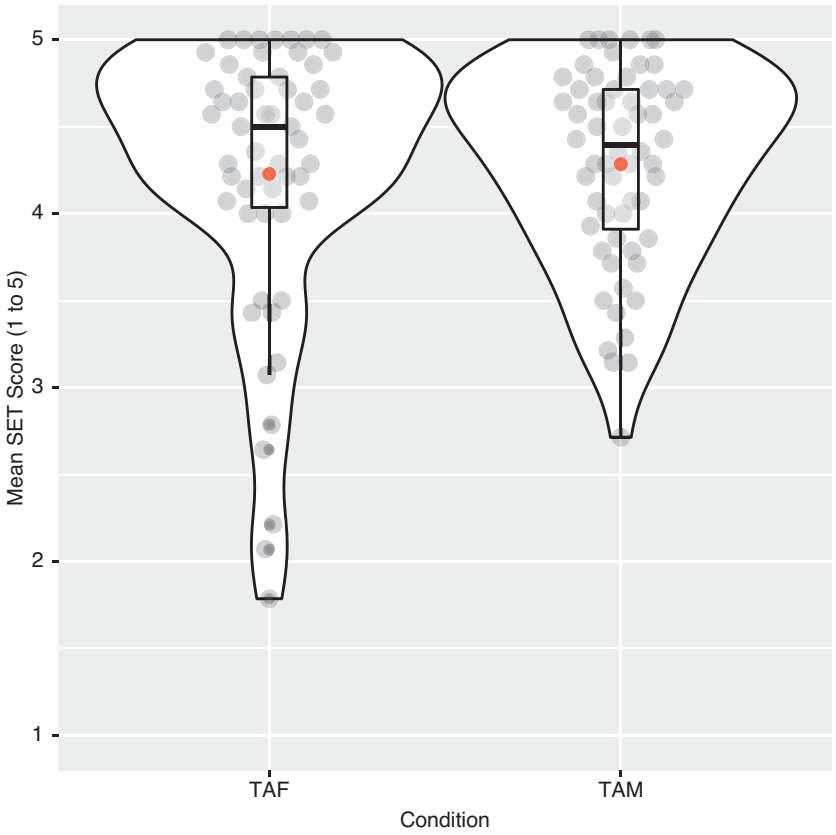

Figure 5. The violin plots of Khazan et al.'s recovered data on the original 1 to 5 Likert SET scale, for TAF and TAM conditions, with the six outliers included. The mean rating of the female TA is slightly lower than that of the male TA but the difference is trivial, 0.05. In contrast, the median rating of the female TA is higher than that of the male TA.

approachable, extraverted/sociable, ostensibly good natured male looking straight into the camera. In contrast, the female photo showed serious, focused, inquisitive female, ignoring the camera, and looking at what appeared to be insects in some netting. Third, only one exemplar female TA – Ms. Khazan – was actually teaching. It is impossible to generalize from this n-equals-one design to other TAs. Fourth, Ms, Khazan was not blind to the condition the students were assigned to, which by itself may have influenced her communication with the students.

Khazan et al.'s [1] study parallels and replicates the findings of the earlier study by MacNell et al. [3]. MacNell et al. examined SET ratings of one female and one male instructor, each teaching two sections of the same online course, and each pretending to be of the opposite gender for one of the two sections. MacNell et al. concluded that their study demonstrated gender bias; they wrote: "Our findings show that the bias we saw here is. actual bias on the part of the students." However, just like Khazan et al., MacNell et al.'s data showed no significant difference between perceived male vs. female instructor overall, that is, on Student Rating Index or average of SET items ($p = 0.128$). MacNell et al. only found statistically significant findings at $p < 0.05$ (i.e., with no adjustment for number of tests) for 3 out of 12 SET items they chose to analyze (out of 14 in total). Moreover, just like Khazan et al., MacNell et al.'s data included three outliers. When the three outliers were removed, MacNell et al. data revealed that students rated the actual female vs. male instructor higher (non-significantly) regardless of perceived gender [5].

Both MacNell et al. and Khazan et al. highlight problems of outliers, of extreme scores, in SETs. Outliers can and do substantially change, typically pull down, the mean SET for a course regardless of who teaches it. Outliers do not discriminate; they pull down the mean SET for female as well as male professors; for TAs as well as for professors; for White, Black, Hispanic, Asians, Indigenous, and other ethnic/racial groups; and for everyone else too. Outliers are a well-known problem that led many universities to abandon means in favour of medians and interpolated medians that are unaffected by outliers, that is, SET ratings that are far removed from the bulk of the distribution.

Khazan et al.'s decision not to disclose their data and subsequent decision to ignore our request to verify the recovered data accuracy is unfortunate; it runs contrary to research transparency and openness in science and undermines credibility of Khazan et al.'s research. It is widely recognized that data sharing is essential in improving research transparency and reproducibility [8]. Many leading journals, for example, Nature (https://www.nature.com/nature-research/editorial-policies/reporting-standards), PLOS

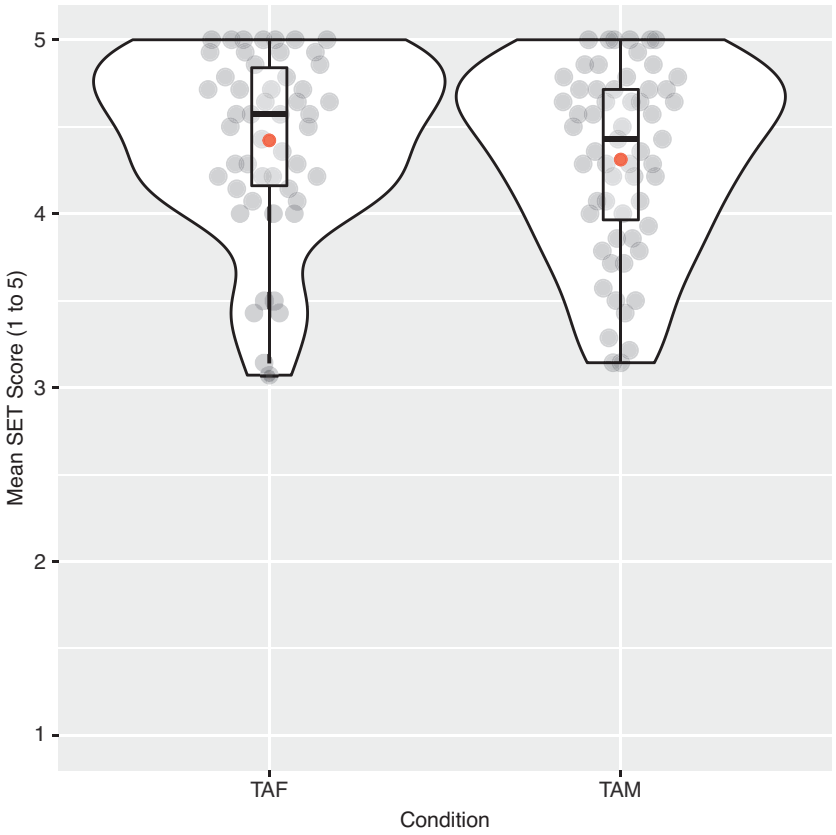

Figure 6. The violin plots of Khazan et al.'s recovered data on the original 1 to 5 Likert SET scale, for TAF and TAM conditions, with the six outliers removed. Both the mean and median ratings for the female TA are higher than for the male TA although not significantly so.

ONE (https://journals.plos.org/plosone/s/data-availability), PeerJ (https://peerj.com/about/policies-and-procedures/#data-materials-sharing), have mandatory data sharing policies. For example, Nature statements says:

An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. A condition of publication in a Nature Research Journal is that **authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications [emphasis in original]**…

Unfortunately, NACTA Journal appears not to have any such policy requiring transparency, openness and data sharing, and Khazan et al. do not appear to endorse such policies in any case. Khazan et al. article should not have been accepted for publication in a scientific journal given the complete disconnect between the authors' claims of gender bias and their data showing no evidence of any such gender bias (p = 0.73). Ms. Khazan's subsequent tweet that Khazan et al. "show gender bias in teaching reviews of graduate students…" is nothing short of astonishing and generated both praise and swift rebuke. However, the disinformation that Khazan et al. found "gender bias" in student evaluation of teaching have been spreading and featured in number of media outlets including University of Florida News, insidehighered.com, wcjt.com, and wiareport.com. Presumably, if the journalists reporting on Khazan et al. study actually read it, they would have realized/read that the mean scores for male vs. female TAs "were not statistically different… p = 0.73…" and would not have participated in spreading this disinformation.

SET are invalid, biased by variety of factors, and ought not to be used to evaluate faculty's teaching effectiveness [9]. SET use violates the basic principles of assessment, various codes of ethics, and human rights codes (Uttl [10]). SET use also harms student learning, destroys academic integrity, and interferes with academic freedom [11, 12]. However, Khazan et al. study provides no evidence that SET are biased by TA gender. Notwithstanding the fatally flawed design, Khazan et al.'s study found no evidence of gender bias in SET ratings and any claims to the contrary is nothing else but a mirage.

ACKNOWLEDGEMENTS

We thank Carrie Leonard for careful reading and comments on the manuscript.

COMPETING INTERESTS

Authors have no conflict of interest to disclose.

REFERENCES

[1] Khazan E, Borden J, Johnson S, Greenhaw L. Examining gender bias in student evaluation of teaching for graduate teaching assistants. *NACTA J*. 2020;64:430–35.

[2] Flaherty C. Gender bias in TA evals. *Inside Higher Ed*. 2020, November 2. Available from: https://www.insidehighered.com/news/2020/11/02/study-finds-gender-bias-ta-evals-too.

[3] MacNell L, Driscoll A, Hunt A. What's in a name: exposing gender bias in student ratings of teaching. *Innov High Educ.* 2015;40(4):291–303. https://doi.org/10.1007/s10755-014-9313-4.

[4] Clark A. Study reveals gender bias in TA evaluations. *University of Florida News*. 2020, November 3. Available from: https://news.ufl.edu/2020/11/ta-bias/.

[5] Uttl B, Violo V.C. 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.

[6] Ioannidis J.P.A. Why most published research findings are false. *PLOS Med.* 2005;2(8):e124. https://doi.org/10.1371/journal.pmed.0020124.

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

[8] Alter G, Gonzalez R. Responsible practices for data sharing. *Am Psychol.* 2018;73(2):146–56. https://doi.org/10.1037/amp0000258.

[9] Uttl B, White CA, Gonzalez DW. Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Stud Educ Evaluation*. 2017;54:22–42. https://doi.org/10.1016/j.stueduc.2016.08.007.

[10] Uttl B. (2021). Lessons learned from research on student evaluation of teaching in higher education. In: Rollett W, Bijlsma H, Röhl S, editors. Student Feedback on Teaching in Schools. Springer, Cham. https://doi.org/10.1007/978-3-030-75150-0_15.

[11] Stroebe W. Why good teaching evaluations may reward bad teaching: on grade inflation and other unintended consequences of student evaluations. *Perspect Psychol Sci*. 2016;11(6):800–16. https://doi.org/10.1177/1745691616650284.

[12] Stroebe W. Student evaluations of teaching encourages poor teaching and contributes to grade inflation: a theoretical and empirical analysis. *Basic Appl Soc Psych.* 2020;42(4):276–94. https://doi.org/10.1080/01973533.2020.1756817.

PUBLISHING NOTES