Double-blind reviewing and gender biases at EvoLang conferences: an update

Abstract

A previous study of reviewing at the Evolution of Language conferences found effects that suggested that gender bias against female authors was alleviated under double-blind review at EvoLang11. We update this analysis in two specific ways. First, we add data from the most recent EvoLang12 conference, providing a comprehensive picture of the conference over five iterations. Like EvoLang11, EvoLang12 used double-blind review, but EvoLang12 showed no significant difference in review scores between genders. We discuss potential explanations for why there was a strong effect in EvoLang11 which is largely absent in EvoLang12. These include testing whether readability differs between genders, though we find no evidence to support this. Although gender differences seem to have declined for EvoLang12, we suggest that double-blind review provides a more equitable evaluation process.

Introduction

The modern academic ecosystem relies heavily on the practice of peer review, from awarding grants to publishing scholarly research (Lee, Sugimoto, Zhang & Cronin, 2012). In this paper, we examine the issue of gender bias in peer review for conference submissions, specifically at the Evolution of Language (EvoLang) conferences, which have occurred every two years since 1998. In particular, we aim to assess how double-blind review (the reviewer is unknown to the authors, and vice versa), as opposed to single-blind review (the reviewer is unknown to the authors, but the authors are known to the reviewer), is associated with different outcomes for male versus female authors submitting to the conference.

Roberts & Verhoef (2016) analysed data from EvoLang 9, 10 and 11 (held in 2012, 2014, and 2016 respectively). EvoLang 11 was the first in the series to introduce double-blind review for all submissions. Roberts & Verhoef (2016) found that for conferences using single-blind review, there was no significant difference in rankings between male and female-authored submissions. However, under double-blind review, the average ranking was higher for submissions by female first authors than by male first authors, and this effect was stronger for senior (non-student) female authors (Roberts et al., under review). This is consistent with the Matilda effect, where the work of female academics is systematically under-valued (Rossiter, 1993; Knobloch-Westerwick, Glynn & Huge, 2013), including in the context of peer review. Roberts & Verhoef argued that their result reflected a bias against female authored work in Language Evolution under single-blind review conditions, and that double-blind review could help mitigate this bias. Note that Roberts & Verhoef (2016) included a minor statistical error, which we have corrected in errata (ANONYMOUS et al., under review).

The current paper extends Roberts and Verhoef (2016) by adding data from EvoLang 8 (2010, single-blind) and the most recent EvoLang 12 conference (2018, double-blind), and

exploring potential mechanisms behind observed gender disparity in rankings (or lack thereof).

The literature on gender biases in scientific review is covered in depth elsewhere (see Roberts & Verhoef, 2016; Snodgrass 2006; Savonick & Davidson, 2017), here, we focus briefly on findings relevant to single- and double-blind review. Historically, single-blind reviewing is the dominant system in place, mainly for administrative reasons: the process of making work anonymous is thought to be more difficult for both authors and journal editors (Lee et al., 2012). One of the criticisms of double-blind reviewing is that this effort is wasted: it does not effectively hide the identity of authors, especially well-known authors. However, Le Goues et al. (2018) found that reviewers are actually fairly poor at guessing author identities in double-blind review: over three computer science conferences, they found 70% of reviewers did not feel confident guessing the identities of authors, and of those that did guess, less than 30% were able to guess all authors correctly.

Single-blind reviewing, on the other hand, may be susceptible to conscious or unconscious bias on the part of reviewers, for example, biases which favour more well-known authors, male authors, or more prestigious institutions. Tomkins, Zhang & Heavlin (2017a) analysed review statistics from the 10th ACM International ACM Conference on Web Search and Data Mining. Each paper was given to two reviewers who were told the names of the authors and two who were not. Compared to double-blind reviewers, single-blind reviewers were more likely to give positive reviews to papers by authors who were well-known in the field or from a top institution. While female authors did receive slightly more negative reviews, this was not statistically significant in their sample.

However, Tomkins, Zhang & Heavlin (2017b) performed a meta-study analysis from 5 other studies (including results from EvoLang 11) and did find a significant overall effect of gender in that female authors were more positively rated under double-blind conditions. In this meta-study, the study of EvoLang 11 showed the strongest bias. Krawczyk & Smyk (2016) ran a controlled experiment which manipulated the gender and age information about the author that was given to reviewers. Gender (but not seniority) biased reviewer's judgements, with articles supposedly written by females being rated as less likely to be accepted.

In contrast, McGillivray & De Ranieri (2018) found no statistically significant difference in the rejection rate between male and female corresponding authors for the two review models (based on a large sample of submissions to Nature journals). While several studies likewise fail to find a statistically significant gender bias in single-blind review (e.g., Blank, 1991, Engqvist & Frommen, 2008; Fox et al., 2016; Handley et al., 2015a), even these studies find that female authors fare measurably better under double-blind conditions (and the significance in some cases is marginal e.g. McGillivray & De Ranieri, $p = 0.054$). This indicates that these differences are unlikely to be the result of random noise: were this the case, we would expect double-blind review to sometimes result in measurably better outcomes for male authors. Given the relevance of gender bias not only to peer review, but across science (Nature Special Issue, 2013) and higher education more broadly (MacNell, Driscoll & Hunt,

2015), extra data is valuable. Therefore, Study 1 extends the timeframe examined by Roberts & Verhoef (2016) by including data from EvoLang 8 (2010) and EvoLang 12 (2018).

The results prompt fuestions about the proximate mechanisms which may underlie any observed differences in rankings between male and female authors. In other words, under double-blind conditions, *what* about female authored abstracts makes them more likely to receive higher ratings? Hengel (2017) analysed journal abstracts in economics and found that those written by women were up to 6% more readable than those written by men. Hengel suggests that "the simplest interpretation is that editors and referees expect clearer, more direct writing from women" (p.1). If female authors do write more clearly than men, this might explain the increase in female scores observed under double-blind review in EvoLang 11. We performed an analysis of readability, but problems ensuring clean transcriptions and a possible lack of power limit the possible insights. The analysis can be found in the supporting materials. Study 2 takes a closer look at the rankings of specific authors under single- versus double-blind conditions, extending similar analyses from Roberts and Verhoef (2016). Lacking a controlled comparison like studies which contrast single and double-blind review of the same work (e.g., Thompkins et al., 2017b), we instead track the ranking of individual authors across single- and double-blind conditions.

Study 1

Data

Review scores were available for conferences 8 to 12 (Smith et al., 2010; Scott-Phillips et al., 2012; Cartmill et al., 2014; Roberts et al., 2016; Cuskley et al. 2018). For each submission, the mean reviewers' score was calculated and the submissions were ranked within each conference based on this mean. We used the ranked scores instead of absolute reviewer scores because mean absolute scores differed significantly between the conferences, likely also as a result of double-blind versus single-blind review. The submission rankings were then scaled within each conference (0 = worst, 1 = best, average rank used for ties). For the statistical modelling, these scaled rankings were then centered and scaled to have a mean of 0 and standard deviation of 1 (see SI). Authors specified their student status for all conferences except EvoLang 8. Gender of the first author was coded (by CC and SR) using a binary male/female categorisation based on a subjective assessment of the authors' performed gender on their academic profile. Authors were assigned anonymous identifiers to track rankings across conferences. The submission format was either a two-page abstract or a six-page paper. Table 1 shows some summary statistics. The mean review rank was 0.5 (min = 0, max = 1, sd = 0.3).

Throughout this paper, only the identity of the first author is considered. Data for 927 submissions were available. Figure 1 shows the number of submissions by gender and student status for each conference. For full data and analysis, see the supporting information or https://github.com/seannyD/EvoLang12DoubleBlindData.

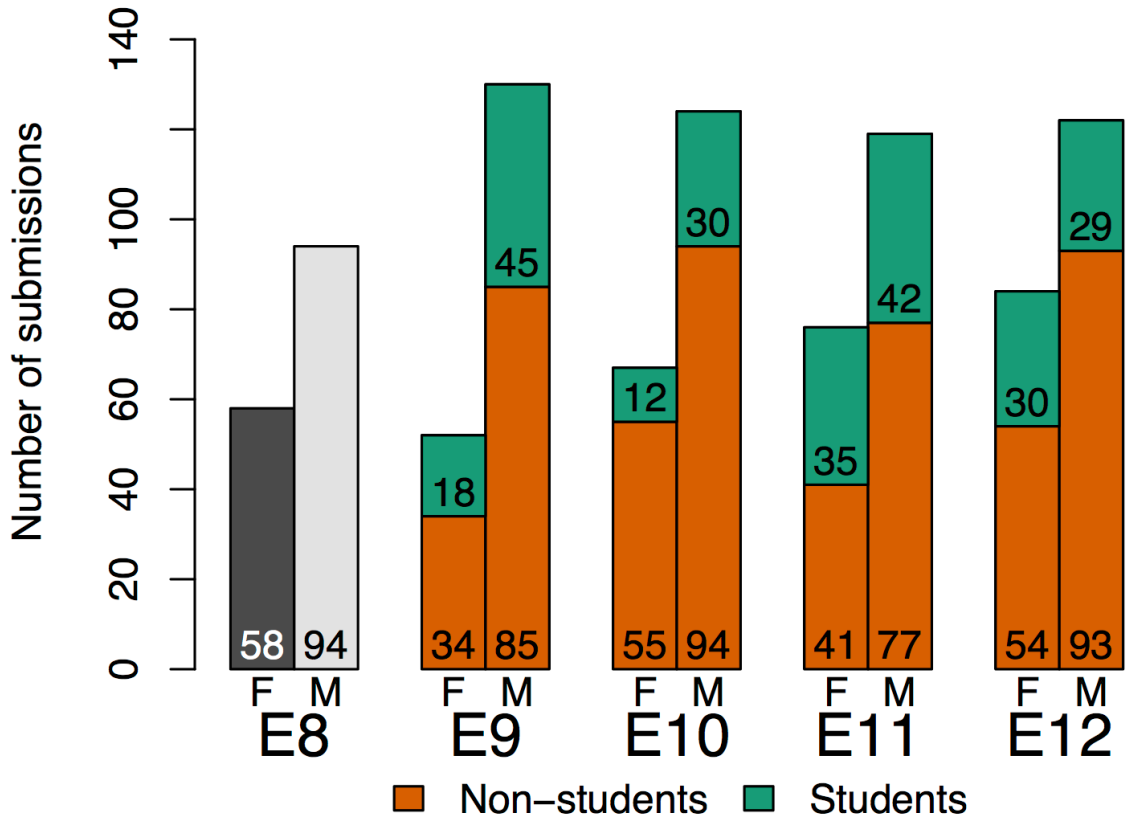

Figure 1: Number of available submissions by conference, gender of first author and student status. Data on student status was not available for EvoLang 8.

	Gender		**Student status**		**Format**	
Conference	Female	Male	Non-Student	Student	Abstract	Paper
E8 (SB)	58	94	NA	NA	98	55
E9 (SB)	52	130	119	63	121	61
E10 (SB)	67	124	149	42	131	60
E11 (DB)	76	119	118	77	145	50
E12 (DB)	84	122	147	59	161	45

Table 1: Summary of the number of papers available in each conference.

Results

Figure 2 shows the distribution of scores by gender for each conference. EvoLang 11 and 12 were conducted with double-blind review. There is a clear difference between genders for EvoLang 11, but the scores for EvoLang 12 look much more even.

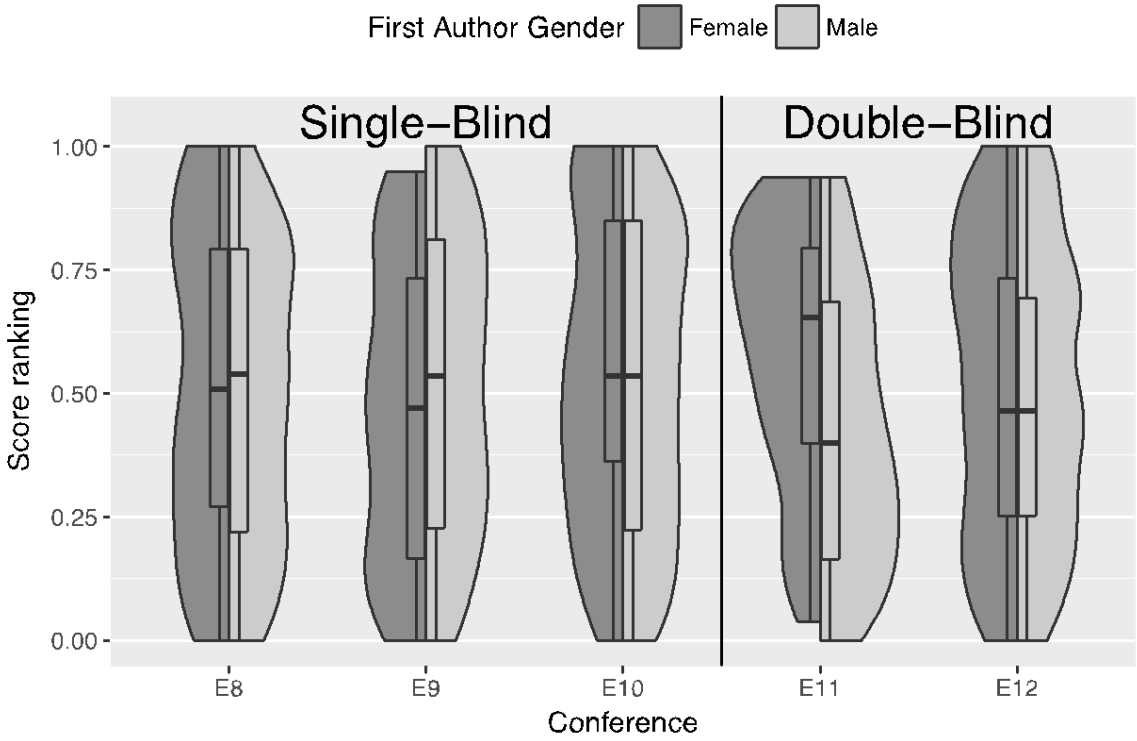

Figure 2: The distribution of review rankings by conference and gender. Conferences are arranged from left to right. For each conference, the distribution is shown for female authors (dark gray, left) and male authors (light gray, right). Each distribution is represented by a box plot (black line = median, box = first and third quartiles, whiskers = 1.5 times the interquartile range), and a violin plot which shows a smoothed density plot of scores.

We performed a four-way independent-samples ANOVA on paper ranking by gender, student status, conference and submission type (abstract or full paper), and all interactions between the independent variables. Since information about student status was not available for EvoLang 8, this was run for conferences 9-12 only, although comparable results are found when omitting student status from the model and analyzing all conferences. There was a significant main effect of gender ($F(1) = 5.7$, $p = 0.017$) and a significant interaction between first author gender and review type ($F(1) = 4.4$, $p = 0.035$). That is, paper ranking was higher for female authors under double-blind review. However, post-hoc t-tests showed that the gender difference was driven almost entirely by the results from EvoLang 11 (E8 $t = 0.6$, $p = 0.55$; E9 $t = -0.87$, $p = 0.39$; E10 $t = 0.75$, $p = 0.45$, E11; $t = 4.4$, $p < 0.0001$; E12 $t = 0.4$, $p = 0.69$). The supplementary materials also show that both a mixed effects model controlling for random effects within each conference, and a permutation

test, come to the same conclusion: there is a significant difference between genders for EvoLang 11, but not for EvoLang 12.

The ANOVA results suggested that there was a significant main effect of submission type ($F(1) = 12.15$, $p < 0.001$). Since EvoLang 10, abstracts have been given higher scores than full papers. However, this was not robust in the mixed effects model ($t = 1.3$, Satterthwaite $p = 0.3$), suggesting that the generalisation does not hold for all conferences. There was also a significant interaction between student status and submission type ($F(1) = 10.4$, $p = 0.001$). For full papers, students are given higher scores than non-students (about 12.9% difference), but for abstracts the difference is very small (students are given slightly lower scores than non-students). This effect was robust in the mixed effects model and in a binary decision tree analysis (using the *party* package in R, Hothorn, Hornik & Zeileis, 2006).

In summary, study 1 found that the effect of double-blind peer-review at EvoLang 11 did not persist significantly at EvoLang 12. The results of EvoLang 11 may have been an anomaly, or caused by some other factor that differs between the conferences (proportion of genders, location, different authors, etc.). Another possibility is that the advantage for female authors in EvoLang 11 occurred because they had better writing (as suggested by Hengel, 2017). Male authors may have changed their strategy after having experienced double-blind review (or they may have read Roberts & Verhoef, 2016; though see Handley et al., 2015b) by investing more effort into writing their submissions for EvoLang 12. We assessed the readability of EvoLang abstracts using methods from Hengel (2017). However, we found no evidence that readability differed by gender, possibly because of the smaller sample size than Hengel (2017), and possibly because of noise in the transcription of submission texts.

Study 2

EvoLang 12 included many authors who had not previously submitted to the conference, which might skew the sample. In order to address the issue of new authors in EvoLang 11, Roberts and Verhoef (2016) identified authors that had submitted to both EvoLang 10 and EvoLang 11, and then analysed the "paired change in ranking". Similarly, study 2 analyses data for authors who submitted to each of the last 3 conferences: EvoLang 10, 11 and 12.

Data

We identified 50 authors who had submitted to EvoLang 10, 11 and 12. Because some authors submitted multiple papers per conference, we only analysed each author's highest ranking paper in each conference.

Results

Figure 3 shows the distribution of review scores by conference and gender for the 50 authors who submitted to EvoLang 10, 11 and 12. As in the original study, there appears to be a change from the single-blind EvoLang 10 to the double-blind EvoLang 11 in that the rankings of female authored increase markedly. However, unlike the data in the full sample above, the distribution for EvoLang 12 looks similar to EvoLang 11. A mixed effects model was fit to the review ranking data, with fixed effects for format type, review condition and gender and with random intercepts for authors (see SI). If female authors fare better under

double-blind conditions than single-blind conditions, we should expect a significant interaction between review type and gender. However, adding this interaction does not significantly improve the fit of the model (log likelihood difference = 1.06, df = 1, $\chi^2 = 2.11$, p = 0.14).

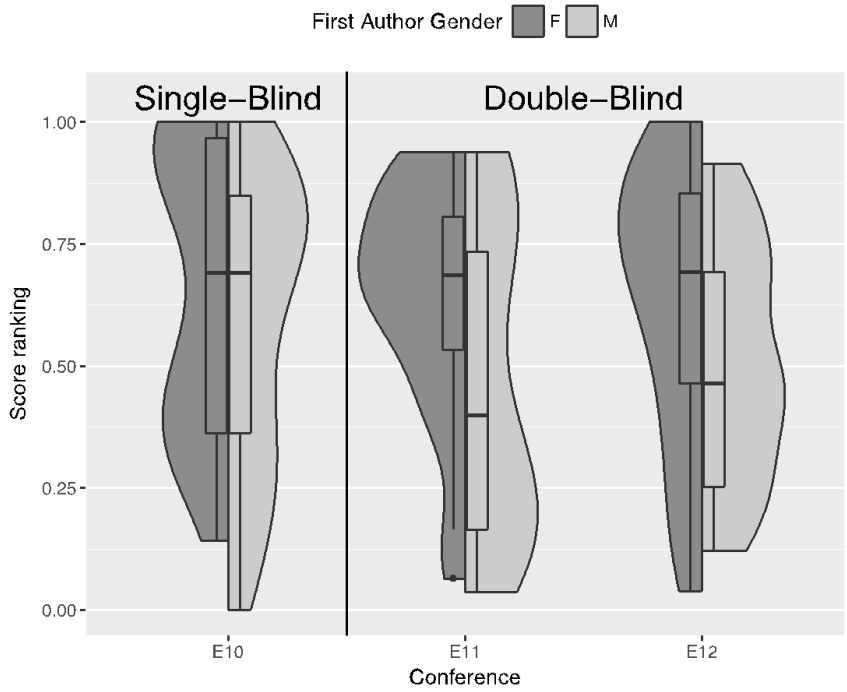

Figure 3: The distribution of average review rankings by conference and gender for 50 authors who submitted to E10, E11 and E12.

Given the complex causal structure of the data, drawing firm conclusions is difficult. For example, the scores are related temporally and submission format varies with gender (male authors are more likely to submit a full paper than female authors). Because abstracts are more likely to be given higher scores than full papers in general, this might confound the effect of gender.

To address this, we fit a structural equation model which accounts for these effects (using the R package *lavaan*, Rosseel 2012, see supporting information). The results agree with the mixed effects model. The direct effect of gender on score for EvoLang 10 is close to zero (β = -0.012, p = 0.93, i.e. the distributions for male and female authors is similar). However, the relationships between gender and score for EvoLang 11 and 12 are positive: in both cases, female authored papers receive higher scores than male authored papers. These relationships fall short of conventional statistical significance thresholds (E11 β = 0.253, p = 0.086; E12 β = 0.274, p = 0.055), but may still warrant consideration given the small sample size. In a more complex model, we found no evidence for student status affecting review scores (see supporting materials).

In summary, when keeping author identities constant, results for the two double-blind conferences seem more similar to each other than in Study 1. This is in line with the original prediction found in Roberts and Verhoef (2016): female authors receive better ranks under double-blind conditions than single-blind conditions. However, as with many earlier studies (e.g., Blank, 1991, Engqvist & Frommen, 2008; Fox et al., 2016; Handley et al., 2015), these

effects are weak or not statistically significant.

Discussion

This paper extended the analysis of gender bias in the Evolution of Language conferences presented in Roberts and Verhoef (2016). Their analyses found that following the introduction of double-blind review at EvoLang 11, female-authored papers were ranked higher on average than in previous conferences. This suggests a potential gender bias in reviewers in previous EvoLang conferences. However, in adding data from the most recent conference, the current paper found that EvoLang 12 did not differ significantly from previous single-blind conferences, despite the review process also being double-blind. However, when authors who submitted to EvoLang 10, 11, and 12 were tracked through the 3 conferences, the patterns present in EvoLang 12 are similar to those found in EvoLang 11: double-blind review leads to higher overall rankings for female authors. However, these patterns fall short of statistical significance. There are a few possible explanations for these findings, which we discuss in detail below.

The first is that the differences observed in EvoLang 11 were an anomaly. The sample sizes in each conference are relatively small, and different conference locations may draw from different authors: EvoLang is historically a predominantly European conference, but EvoLang 11 took place in the US. Additionally, there may be other factors of author characteristics that were not taken into account. As Webb, O'Hara & Freckleton (2008) suggest for other studies, the observed differences may just be random fluctuations over time or driven by different proportions of female and male submissions. Our study does not involve the ideal control of single-blind review scores for EvoLang 11 and 12 (or double blind reviewer for earlier conferences), making direct comparison more complicated.

Another possible explanation is that authors changed their submission strategies after becoming more aware of possible gender biases. This may be due to their previous experience with double-blind review in EvoLang 11 or more general gender bias issues being increasingly visible in media discourse in the last two years (Park, 2017; Bell & Koenig, 2017; Yammine et al. 2018; Nauska, 2018). For example, if male authors generally put less effort into writing submissions (Hengel, 2017), then they could have reacted to the difference in EvoLang 11 by increasing their effort. The apparent gender bias in single-blind reviewing for conferences prior to EvoLang 11 was published in Roberts and Verhoef (2016) and summarised at the general meeting following EvoLang 11, making this result relatively well-known among likely submitees to the conference.

There are many indicators of a recent increase in general awareness of gender-related issues in cognitive science; for example, following the lead of many other conferences, EvoLang 12 instantiated a code of conduct which explicitly mentions gender biases (http://evolang.org/torun/proceedings/conduct.html). Moreover, recent work on 'tipping points' suggests that change in behaviour can happen quite quickly once a critical mass is achieved (e.g. Centola et al., 2018). Nonetheless, we cannot conclude that centuries of bias were removed in just one conference.

This study considered first authors, but future research could explore the effect of supervising authors and institutions. We include counts of papers with multiple authors by conference and gender in table 2. We note that there are generally a higher proportion of male senior authors, and there is a bias for male first authors to work with male last authors. This might be due to female scholars leaving (or being driven out of) academia at a higher rate than males (e.g. Shaik & Fusulier, 2015; Sleeman, Koffman & Higginson, 2017), meaning women are generally under-represented at more senior levels. However, we found no statistical relationship between last author gender and review scores (see supporting materials). Having said this, the data in this study is not ideal for exploring this issue, since the number of papers with multiple authors varies between conferences and there are many non-independencies.

Conference / Last Author		E9 F	E9 M	E10 F	E10 M	E11 F	E11 M	E12 F	E12 M	Total
First Author	F	9	18	8	23	8	19	17	37	139
	M	1	42	13	38	9	29	18	53	203
Total		10	60	21	61	17	48	35	90	342

Table 2: Papers with multiple authors by conference and gender.

Conclusions

In summary, there are several potential explanations as to why EvoLang 11 showed a significant difference between ratings for male and female authors, while this difference failed to reach significance in EvoLang 12. It may be that EvoLang 11 (or EvoLang 12) was an anomaly. As with earlier studies, it may be that the observed differences simply do not reach significance in such a small sample. It is also possible that the publicised differences observed in EvoLang 11, combined with a marked increase in general awareness of issues surrounding gender bias, led to a rapid reconfiguration of submission strategies. However, such rapid change in the community seems unlikely. Further study of trends at future iterations of EvoLang, and of gender bias in double and single-blind reviewing more generally, remains essential.

Despite the fact that EvoLang 12 did not show the strong gender differences found in EvoLang 11, we do not conclude that double-blind reviewing is ineffective, or that a return to single-blind review would be warranted. There is general support for double-blind review reducing various kinds of bias in addition to gender (e.g. Budden et al., 2008; Snodgrass, 2006; Seeber & Bacchelli, 2017). Moreover, there is little evidence for strong *disadvantages* to double-blind review (though see e.g. Schulzrinne, 2009, Tricco et al., 2018). Indeed, the widespread adoption of single-blind review seems to have emerged largely due to perceived administrative burden, rather than for any substantive scholarly reason (Lee et al., 2012). For EvoLang, this burden has been negligible on the editorial side, and appears not to have affected the attitudes of authors - indeed, submission rates have been steadily increasing. The current equality in ratings is promising, especially alongside the increasing number of submissions by female researchers. We aim to continue to collect data on this issue particularly in the EvoLang community as part of a larger effort to monitor our biases as researchers.

Acknowledgements

Added after review.

References

Bell, R. E., & Koenig, L. S. (2017). Harassment in science is real. *Science* 358(6368), pp. 1223.

Blank, R. M. (1991). The effects of double-blind versus single-blind reviewing: Experimental evidence from the American Economic Review. *The American Economic Review*, 1041-1067.

Budden, A. E., Tregenza, T., Aarssen, L. W., Koricheva, J., Leimu, R., & Lortie, C. J. (2008). Double-blind review favours increased representation of female authors. *Trends in ecology & evolution*, *23*(1), 4-6.

Cartmill, E. A., Roberts, S. G., Lyn, H. & Cornish, H. (2014). The Evolution of Language: Proceedings of the 10th International Conference (EVOLANG10).

Centola, D., Becker, J., Brackbill, D., & Baronchelli, A. (2018). Experimental evidence for tipping points in social convention. *Science*, *360*(6393), 1116-1119.

Cuskley, C., Flaherty, M., Little, H., McCrohon, L., Ravignani, A. & Verhoef, T. (2018): The Evolution of Language: Proceedings of the 12th International Conference (EVOLANG12). doi:10.12775/3991-1.099

Engqvist, L., & Frommen, J. G. (2008). Double-blind peer review and gender publication bias. *Animal Behaviour*, *76*(3).

Fox, C. W., Burns, C. S., Muncy, A. D., & Meyer, J. A. (2016). Gender differences in patterns of authorship do not affect peer review outcomes at an ecology journal. *Functional ecology*, *30*(1), 126-139.

Le Goues, C. L., Brun, Y., Apel, S., Berger, E., Khurshid, S., & Smaragdakis, Y. (2018). Effectiveness of anonymization in double-blind review. *Communications of the ACM*, *61*(6), 30-33.

Handley, G., Frantz, C.M., Kocovsky, P.M., DeVries, D.R., Cooke, S.J. & Claussen, J. (2015a) An Examination of Gender Differences in the American Fisheries Society Peer-Review Process, Fisheries, 40:9, 442-451

Handley, I. M., Brown, E. R., Moss-Racusin, C. A., & Smith, J. L. (2015b). Quality of evidence revealing subtle gender biases in science is in the eye of the beholder. *Proceedings of the National Academy of Sciences*, *112*(43), 13201-13206.

Hengel, E. (2017). Publishing while Female. Are women held to higher standards? Evidence from peer review. *Cambridge Working Papers in Economics* 1753, Faculty of Economics, University of Cambridge. https://ideas.repec.org/p/cam/camdae/1753.html

Knobloch-Westerwick, S., Glynn, C. J., & Huge, M. (2013). The Matilda effect in science communication: an experiment on gender bias in publication quality perceptions and collaboration interest. *Science Communication*, *35*(5), 603-625.

Krawczyk, M., & Smyk, M. (2016). Author' s gender affects rating of academic articles: Evidence from an incentivized, deception-free laboratory experiment. *European Economic Review*, *90*, 326-335.

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2-17.

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.

McGillivray, B., & De Ranieri, E. (2018). Uptake and outcome of manuscripts in Nature journals by review model and author characteristics. *arXiv preprint arXiv:1802.02188*. See https://peerreviewcongress.org/prc17-0305.

Nature Special Issue (2013) Women in Science, *Nature* , 495/7439: 5–134.

Nauska, K. (2018) *The Impact of the #MeToo Phenomenon on Working Conditions for Women in Finland: Can Social Media Pressure Bring Change in Gender Politics?* BA thesis. Helsinki Metropolia University of Applied Sciences. https://www.theseus.fi/bitstream/handle/10024/148566/Nauska%20Kaisa.pdf?sequence=1

Park, A., 2017. CBS News. #MeToo reaches 85 countries with 1.7M tweets. Available at: https://www.cbsnews.com/news/metoo-reaches-85-countries-with-1-7-million-tweets/

Roberts, S. G., & Verhoef, T. (2016). Double-blind reviewing at EvoLang 11 reveals gender bias. *Journal of Language Evolution*, *1*(2), 163-167.

Roberts, S., Cuskley, C., McCrohon, L., Barceló-Coblijn, L., Feher, O., & Verhoef, T. (2016). The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11).

Rossiter, M. W. (1993). The Matthew Matilda Effect in Science. Social Studies of Science, 23(2), 325–341. https://doi.org/10.1177/030631293023002004

Yves Rosseel (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1-36.

Savonick, Danica and Davidson, Cathy, (2017) Gender Bias in Academe: An Annotated Bibliography of Important Recent Studies.CUNY Academic Works. http://academicworks.cuny.edu/qc_pubs/163

Schulzrinne, H. (2009). Double-blind reviewing: more placebo than miracle cure?. *ACM SIGCOMM Computer Communication Review*, *39*(2), 56-59.

Scott-Phillips, T., Tamariz, M., Cartmill, E. A. & Hurford, J. R. (2012). The Evolution of Language: Proceedings of the 9th International Conference (EVOLANG9).

Seeber, M., & Bacchelli, A. (2017). Does single blind peer review hinder newcomers?. *Scientometrics*, *113*(1), 567-585.

Smith, A., Smith, K., Schouwstra, M & de Boer, B. (2010). The Evolution of Language: Proceedings of the 8th International Conference (EVOLANG8).

Snodgrass, R. (2006). Single-versus double-blind reviewing: an analysis of the literature. *ACM Sigmod Record*, *35*(3), 8-21.

Tomkins A, Zhang M, Heavlin WD (2017b) Single versus double blind reviewing at WSDM 2017. arXiv:1702.00502.

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017a). Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, *114*(48), 12708-12713.

Torsten Hothorn, Kurt Hornik and Achim Zeileis (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15(3), 651--674.

Tricco, A. C., Thomas, S. M., Antony, J., Rios, P., Robson, R., Pattani, R., Ghassemi, M., Sullivan, S., Selvaratnam, I., Tannenbaum, C. & Straus, S. E. (2017). Strategies to prevent or reduce gender bias in peer review of research grants: a rapid scoping review. *PloS one*, *12*(1), e0169718.

Webb, T. J., O'Hara, B., & Freckleton, R. P. (2008). Does double-blind review benefit female authors?. *Trends in ecology & evolution*, *23*(7), 351-35.

Yammine, S. Z., Liu, C., Jarreau, P. B., & Coe, I. R. (2018). Social media for social change in science. *Science*, *360*(6385), 162-163.