**Six Things Peer Reviewers Can Do To Improve Political Science**

Ideally, published political science would provide high-quality information to policymakers and the public. One of the most important mechanisms for increasing the quality of published political science is peer review, in which qualified persons evaluate a manuscript before publication. Peer review received widespread support in a survey of APSA members (Djupe 2015). However, the peer review process often misses major errors and shortcomings in submitted manuscripts (Baxt et al. 1998, Godley et al. 1998, Schroter et al. 2008). The present manuscript supplements prior advice on peer reviewing (e.g., Miller et al. 2013, Nyhan 2015) to suggest six actions that peer reviewers of quantitative political science can take to improve the representativeness and validity of political science publications, and, for each recommended action, illustrates how the recommended action could have improved published political science.

**To increase representativeness**

Political science that is published would ideally be representative of the political science that has been conducted; otherwise, non-representative publication could produce research that does not replicate, if, for instance, researchers selectively report only studies or models for which their hypotheses have been confirmed. The opportunity for political science estimates and inferences to be non-representative is substantial: measured associations within a study can vary based on factors such as estimation technique, model specification, and inclusion criteria, and measured associations across studies can vary based on factors such as random assignment error, sampling error, and measurement error. Researchers thus can possess flexibility to selectively report from a range of studies and from a range of associations in a given study, which can in turn produce misleading inferences about the replicability of published associations. However, peer reviewers can help increase the representativeness of published political science with the following three actions:

1. Request reporting of novel alternate specifications

…

2. Request declarations about related research

…

3. Check for available relevant information on the research

Peer reviewers do not need to rely only on author disclosures, when relevant information is publicly available...

For another example, consider Study 3 of Valentino et al. (2018), based on data collected from Amazon's Mechanical Turk in which "subjects were randomly assigned to one of three conditions, each inducing a distinct emotional state—anger, fear, and relaxation (the control condition)" (p. 227). Results for this Mechanical Turk study closely correspond to results for a "Pilot Mechanical Turk Study" that was described in a funding proposal (Wayne 2016) for one of the Valentino et al. (2018) studies, with the proposal available at the Time-sharing Experiments for the Social Sciences repository. Footnote 2 of the proposal indicated that the Pilot Mechanical Turk Study included a disgust condition that provided "no empirical support" for a hypothesized relationship between induced disgust, sexism, and support for Donald Trump. Had Valentino et al. (2018) reflected a peer reviewer or editor request for information about this Pilot Mechanical Turk Study, the article might have indicated whether the Mechanical Turk study reported on in the article was the same study as the Pilot Mechanical Turk Study and, if so, might have included the hypothesis and results for this disgust condition.

**To increase validity**

Research results that are fully reported and replicable can nonetheless produce misleading inferences if the underlying research methods are not valid. Peer reviewers can help increase the validity of results reported in published political science with the following three actions:

4. Request tests of discriminant validity

…

5. Request reporting of summary statistics for regression-based inferences about groups

Regression results in isolation can produce misleading or incomplete inferences. Consider this measure of sexism used in a sexism battery reported on in Valentino et al. (2018): "Should the news media pay more attention to discrimination against women, less attention, or the same amount of attention they have been paying lately?". If, in an election between Candidate A and Candidate B, each Candidate A voter reported that the news media should pay a great deal more attention to discrimination against women and each Candidate B voter reported that the news media should pay somewhat more attention to discrimination against women, then in a sufficiently large sample a regression would permit the inference that sexism predicted vote choice for Candidate B, even though Candidate B voters reported the non-sexist desire for the news media to pay more attention to discrimination against women. Regression models would focus attention on only the relative position of Candidate A voters and Candidate B voters, but the absolute location of Candidate A voters and Candidate B voters on a sexism scale is an important consideration when assessing the role of sexism in vote choice.

For an example of how such a peer review request might work, Figure 1 reports mean responses for Hillary Clinton voters and for Donald Trump voters on the three items used in the sexism measure in Study 2 of Valentino et al. (2018), based on data from the American National Election Studies 2016 Time Series Study (ANES 2017). The mean response from Trump voters reflects these beliefs: the news media should pay the same amount of attention to discrimination against women as has been paid lately; when women complain about discrimination, less than half of the time they cause more problems than they solve; and, when women demand equality these days, less than half of the time they are actually seeking special favors. The validity of the Valentino et al. (2018) inference that "sexism was significantly associated with voting for the Republican candidate" (p. 225) could have been strengthened had the manuscript authors followed a peer reviewer request to include these summary statistics and explain why the mean reported beliefs of Trump voters on these scales should be considered sexist.

[Figure 1 about here]

6. Request a statistical test for inferences about group differences

Political scientists sometimes hypothesize that the association between two variables will differ between years or groups or other phenomena, such as "compared to recent elections, the impact of sexism should be larger in 2016…" (Valentino et al. 2018: 219-220), and… However, neither of these cited publications hypothesizing such a difference reported a statistical test of whether the coefficients on their sexism measures differed as hypothesized (cf. the first item in the Nyhan 2015 peer review checklist).[[1]](#footnote-1)

…

**Discussion**

…



**Figure 1. Mean responses on Valentino et al. (2018) Study 2 sexism items for the 2016 U.S. presidential election**

Note: The figure reports mean weighted responses for Hillary Clinton voters and Donald Trump voters on the three modern sexism items used in Study 2 of Valentino et al. (2018) for assessing sexism in the 2016 U.S. presidential election vote choice. Data source: ANES 2016 Time Series Study.

1. Both publications noted the presence of a difference in statistical significance between certain models (Valentino et al. 2018: 225; ...). However, as Gelman and Stern (2006) noted, "comparisons of the sort, 'X is statistically significant but Y is not,' can be misleading" (p. 331). For example, in Valentino et al. (2018), the sexism predictor in the 2016 model had a logit coefficient and standard error of 1.69 and 0.81. The statistical significance of this sexism predictor indicates that there is sufficient evidence at p<0.05 that the sexism coefficient differs from zero, but the statistical significance of this sexism predictor does not indicate whether that there is sufficient evidence at p<0.05 that the sexism coefficient differs from the imprecisely estimated sexism coefficients for the 2012, 2008, and 2004 models of 0.23, 0.94, and 0.34. [↑](#footnote-ref-1)