CAMBRIDGE
UNIVERSITY PRESS

REPLICATION STUDY 📊 ⬡ ✅

Does Stereotype Threat Contribute to the Political Knowledge Gender Gap?
A Preregistered Replication Study of Ihme and Tausendpfund (2018)

Flavio Azevedo[1]* 🆔, Leticia Micheli[2] 🆔 and Deliah Sarah Bolesta[3] 🆔

[1]Department of Psychology, University of Cambridge, Cambridge, UK, [2]Institute of Psychology, Leiden University, Leiden, The Netherlands and [3]Center for Criminological Research Saxony (ZKFS), Technical University Chemnitz, Chemnitz, Germany
*Corresponding author. Email: fa441@cam.ac.uk

Abstract
The gender gap in political knowledge is a well-established finding in Political Science. One explanation for gender differences in political knowledge is the activation of negative stereotypes about women. As part of the Systematizing Confidence in Open Research and Evidence (SCORE) program, we conducted a two-stage preregistered and high-powered direct replication of Study 2 of Ihme and Tausendpfund (2018). While we successfully replicated the gender gap in political knowledge – such that male participants performed better than female participants – both the first ($N = 671$) and second stage ($N = 831$) of the replication of the stereotype activation effect were unsuccessful. Taken together (pooled $N = 1,502$), results indicate evidence of absence of the effect of stereotype activation on gender differences in political knowledge. We discuss potential explanations for these findings and put forward evidence that the gender gap in political knowledge might be an artifact of how knowledge is measured.

Keywords: Stereotype threat; political knowledge; gender gap; political behavior; replication; open data; preregistered direct replication

The gender gap in political knowledge is considered "one of the most robust findings in the field of political behavior" (Dow 2009, 117) and is thought to be linked to women's lower political participation and representation (Ondercin and Jones-White 2011). The underlying reasons for this knowledge gap, however, remain

contentious. Until recently, most studies focused on cultural and macro-level factors (Burns, Schlozman and Verba 2001; Carpini and Keeter 2005). In contrast, Ihme and Tausendpfund (2018) offered a psychological explanation. Specifically, they explored whether the activation of negative stereotypes about women's lower political knowledge can harm women's performance.

According to the stereotype threat literature, exposure to negative stereotypes about one's in-group increases anxiety, negative thinking, and psychological discomfort, all of which overload the working memory and ultimately hamper cognitive performance (McGlone and Pfiester 2007; Pennington, Heim, Levy, and Larkin 2016). These psychological processes, in turn, reinforce the existing stereotypes (Schmader, Johns, and Forbes 2008). Conversely, non-stigmatized individuals exhibit an enhanced task performance when exposed to negative stereotypes related to their outgroup (i.e., a "stereotype lift"; Walton and Cohen, 2003). Consistent with both stereotype threat and stereotype lift, Ihme and Tausendpfund (2018) found that female participants performed worse than males in a political knowledge test when gender stereotypes were activated ($N = 377$). They also observed no knowledge gap in the absence of activated gender stereotypes. Specifically, women performed worse (and men performed better) when gender stereotypes were activated compared to a control condition. These findings persisted even when controlling for political interest, ruling out that the results are a function of women's lack of interest on the topic. Further, the effect of stereotype threat on the gender gap in political knowledge was more pronounced for female students of Politics, presumably because the test represented higher stakes for them as supposedly experts on the topic. The authors concluded that "the often-found gender gap in political knowledge might – to some extent – be the result of stereotyping" (Ihme and Tausendpfund 2018, 12). These findings represent an important practical contribution, as they suggest that the political knowledge gender gap is not necessarily stable, and thus could be potentially mitigated by a range of interventions (for a review, see Lewis and Sekaquaptewa 2016).

The effects of stereotype threat on gender differences in performance have not been consistent in the literature. Pruysers and Blais (2014) found no effect of stereotype threat on the political knowledge gap. McGlone, Aronson and Kobrynowicz (2006) found that implicit and explicit cues of gender stereotype threats impaired women's performance on a political knowledge test, but did not improve males' performance. Adding to the contention, careful examinations of stereotype threat effects on other domains, such as women's and girls' mathematics performance, reveal at most weak evidence in its favor (Flore and Wicherts 2015; Flore, Mulder and Wicherts 2018; Pennington, Litchfield, McLatchie, and Heim 2018). These inconsistent patterns call into question whether the effect of stereotype threat on the political knowledge gap is replicable and, if so, to what extent. To date, no direct replication of this effect has been conducted.

As part of a large-scale replication initiative led by the Center for Open Science and SCORE program (Systematizing Confidence in Open Research and Evidence; https://www.cos.io/score), aiming to investigate the credibility of scientific claims in social and behavioral sciences (Alipourfard et al. 2021), we have conducted a preregistered (peer-reviewed), well-powered, two-step direct replication of Ihme and Tausendpfund (2018).

Methods

As determined by SCORE, the focal claim we attempted to replicate was that "the activation of gender stereotypes affects performance on a political knowledge test" (Ihme and Tausendpfund 2018, 1). As in the original study, we employed a 2 (gender: male vs. female) × 2 (field of study/work: non-politics vs. politics) × 3 (stereotype activation: stereotype activated by gender question vs. stereotype activated by gender difference statement vs. stereotype not activated) between-subjects design. Note that the original study included the variable field of study in all reported analyses. Thus, even though this variable was not necessary to the replication of the effect of gender stereotype activation on political knowledge, we included it in our direct replication so our study design and analyses were as similar and comparable as possible to the original study. According to SCORE guidelines, the replication would be deemed successful if the statistical results showed a significant interaction ($\alpha = 0.05$) between stereotype activation and gender. All study materials, containing ethical approval, power calculation, and preregistration, are publicly available at OSF (https://osf.io/8feku/?view_only=99a41a96c8cd43c4 ab349e44d79919cd).

Sample

The required sample size for replicating the focal claim was determined with power analyses carried out using the "pwr" package (Champely 2020) in R (R Core Team 2020). Power calculations were performed in accordance with the guidelines of the Social Sciences Replication Project (http://www.socialsciencesreplicationproject. com/). As per SCORE guidelines, data collection should proceed in two stages, with a second round of data being collected only if the first round resulted in an unsuccessful replication. Two power calculations were then performed to derive the sample sizes required for each stage of data collection. For the first round of data collection, 90% of power should be achieved. Assuming that the true effect size of the interaction term between gender and stereotype activation was 75% of that reported in the original study, the power analysis yielded a sample of 667 participants. The pooled sample (including both the first and second stages of data collection) should achieve 90% power. Assuming that the true effect size of the interaction between gender and stereotype activation was 50% of that reported in the original study, the second power analysis suggested an additional 830 responses would be needed. Participants were recruited using a professional survey firm (https://www.cint.com) using attention checks as recommended (Aronow, Kalla, Orr, and Ternovski 2020). Only American citizens older than 18 years studying or working at the time of the survey were invited to take part.

Procedure

To ensure a fair and reliable replication attempt, the study design and analysis plan were peer-reviewed by independent researchers selected by SCORE and preregistered on OSF (https://osf.io/nxrg7). The study was approved by an independent IRB ethics committee, BRANY (https://www.brany.com), and the U.S. Army's

Human Research Protection Office (HRPO)#20-032-764 (Award Number HR00112020015, HRPO Log Number A-21036.50).

According to existing definition efforts (Parsons et al. 2021), our study can be considered a direct replication of Study 2 by Ihme and Tausendpfund (2018) as it uses the same methodology and experimental design employed by the authors of the original study, with few modifications as follows. First, our sample was composed not only of students, as in the original study, but also of working adults. This modification was necessary to achieve the required sample size, which was considerably higher than the original study, and to check whether the original findings (in German students) generalize to the adult population of the United States. As a consequence, the political knowledge scale used in our study had to be adapted from a German political scenario to the contemporary political context of the United States (see Table S1). Second, as our sample was composed of both students and working adults, the measurement of participants' field of study had to be expanded to encompass fields of study or work. Data were collected online and hosted at Qualtrics. Both stages of data collection had exactly the same procedures and measures. Before participants answered the political knowledge test, we measured *political interest* and manipulated *stereotype activation* in the same way as Ihme and Tausendpfund (2018). We provide additional sample, procedural, and question wording details in the Supplementary Materials.

Data analysis

Following the analyses reported in the original study and the analysis script made available by the original authors, we tested the replication claim that activation of gender stereotypes influences performance in a political knowledge test with a 2 (gender) × 2 (field of work/study) × 3 (Gender Stereotype Activation) ANCOVA. The dependent variable was participants' total score on the political knowledge test. As in the original study, a single score of political interest was calculated per participant (i.e., average of responses in the short scale of political interest) and included as a covariate. In addition, we use Bayesian analyses to adjudicate about whether results indicate absence of evidence or evidence of absence. All analyses were conducted in R. To increase comparability between the direct replication and original results, we adjusted the sum of squares in R to type III, which is the default in the SPSS software used by the original authors to perform their analyses.

Results

Stage 1

Results of the ANCOVA yielded a non-significant interaction between stereotype activation and gender, $F(2, 658) = 0.691$, $p = 0.501$, partial $\eta^2 = 0.002$, 95% CI = [.00, .01], $N = 671$. Thus, according to the SCORE criteria, the replication was considered unsuccessful at the first stage (see Tables S2–S4 for detailed results). As preregistered, to provide further evidence regarding the (non)replicability of gender stereotype threat on gender differences in political knowledge, we then proceeded to a second stage of data collection.
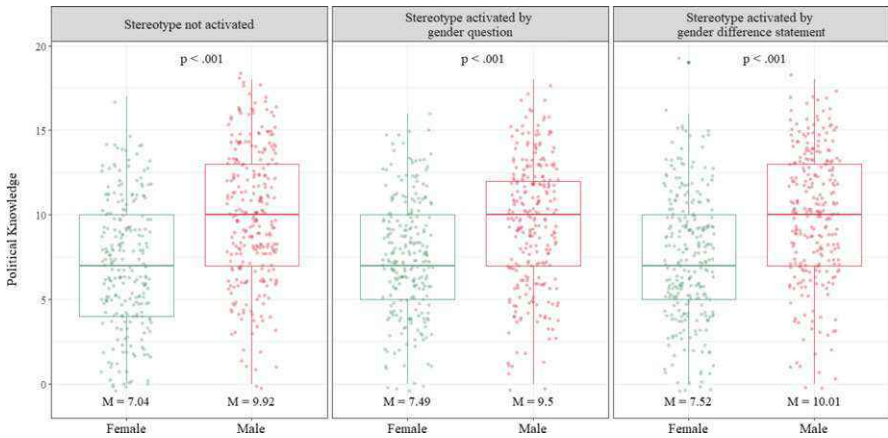
Figure 1

Unconditional means comparison of Political Knowledge Test scores for each gender and experimental condition.

Stage 2

The pooled analytical sample (first and second stages together) was composed of 1,502 participants ($M_{age} = 45.87$ years, $SD_{age} = 17.35$, 48.74% female). The distribution of participants across conditions resembled the distribution of the original study (see Table S5). Consistent with the original study and a large body of research, ANCOVA results revealed a main effect of gender on political knowledge, such that men generally scored higher than women on the political knowledge test, $F(1, 1489) = 28.61$, $p < 0.001$, partial $\eta^2 = 0.02$, 95% CI $= [.01, .04]$; $M_{female} = 7.36$, $SD = 3.62$; $M_{male} = 9.81$, $SD = 3.87$. Also in line with the original study, we found no main effect of stereotype activation on political knowledge, $F(2, 1489) = 0.27$, $p = 0.77$, partial $\eta^2 = 0.00$, 95% CI $= [.00, .00]$, and a significant effect of political interest, such that the more participants were interested in politics, the higher their score on the political knowledge test $F(1, 1489) = 194.78$, $p < 0.001$, partial $\eta^2 = 0.12$; 95% CI $= [.09, .15]$. Our focal test, however, diverged from the results reported in the original study, as the interaction between gender and stereotype activation was not significant $F(2, 1489) = 1.22$, $p = 0.3$, partial $\eta^2 = 0.00$, 95% CI $= [.00, .01]$. Thus, according to the criteria outlined by SCORE, the replication of the effect of stereotype threat on the gender gap on political knowledge was unsuccessful even after the second stage of data collection.

We further explore the results by conducting Bonferroni-corrected pairwise comparisons with the *emmeans* function in R (Lenth 2022). As illustrated in Figure 1, males' scores were significantly higher than females' in the stereotype not activated condition $t(1489) = -7.42$, $p < 0.001$, the stereotype activated by gender question $t(1489) = -4.36$, $p < 0.001$, and in the gender difference statement condition $t(1489) = -6.02$, $p < 0.001$. In addition, we did not find evidence of either stereotype threat or stereotype lift, as women's performance did not decrease nor men's performance increased in the stereotype-activated conditions compared to the stereotype not activated condition (see supplementary materials section 3.3

for detailed analyses). The interaction between field of study/work and stereotype activation as well as the three-way interaction between field of study/work, stereotype activation, and gender were not significant ($p = 0.32$ and $p = 0.81$, respectively). Additional analyses and a comparison between the replication results and the results of the original study can be found in the supplementary materials (Tables S6–S7).

Exploratory analyses

In order to evaluate our replication attempt, we computed the *evidence-updated replication Bayes factors* for both stages of data collection (Ly, Etz, Marsman, and Wagenmakers 2019; Verhagen and Wagenmakers 2014). Using the "posterior distribution obtained from the original study as a prior distribution for the test of the data from the replication study" (Ly, Etz, Marsman, and Wagenmakers 2019, 2504), we computed an overall Bayes Factor of $BF_{10}(d_{orig}, d_{rep}) = 0.009$ for the interaction term of gender and stereotype activation on political knowledge at Stage 1. Dividing the overall Bayes factor by the Bayes factor from the original data ($BF_{10}(d_{orig}) = 0.142$) yielded a replication Bayes factor of $BF_{10}(d_{orig} \mid d_{rep}) = 0.064$. For Stage 2, an overall Bayes Factor of $BF_{10}(d_{orig}, d_{rep}) = 0.001$ for the interaction effect of gender and stereotype activation on political knowledge was computed. Again, dividing this by the original study's Bayes Factor resulted in a replication Bayes factor of $BF_{10}(d_{orig} \mid d_{rep}) = 0.007$. This means that the replication data are predicted $1/0.064 = 15.8$ (Stage 1) or $1/0.007 = 143$ (Stage 2) times better by the null hypothesis than by the alternative hypothesis in the original dataset. Hence, the replication cannot be deemed successful (Zwaan, Etz, Lucas, and Donnellan 2018).

In addition, we evaluated – as per original authors' advice – whether the political knowledge scale is a "sufficiently difficult test." Using Ihme and Tausendpfund (2018) original data, we compared scales' difficulty. Comparing the political knowledge test distribution of the original and the replication data revealed no significant differences for Stage 1 ($z = -1.53$, $p = 0.06$) or Stage 2 ($z = -1.22$, $p = 0.11$, see Figure 2). To test this more in depth, we used Item Response Theory two-parameter model (2PL). As indicated in Figure 3, both scales display equivalent levels of reliability across the latent construct θ (panel a), show equivalent test difficulty and total score across θ levels (panel b), and – albeit some inter-item differences – have overall corresponding item difficulties (panel c). These findings suggest comparable scale properties for both the original and replication, allowing us to rule out measurement-related (difficulty) issues underlying the non-replication. A variety of robustness checks and additional exploratory analyses are reported in the Supplementary Materials (Tables S8–S20).

Discussion

Ihme and Tausendpfund (2018) have proposed that the activation of negative gender stereotypes accounts for the variance of the political knowledge gender gap. In our independent and well-powered direct replication, we find no evidence that
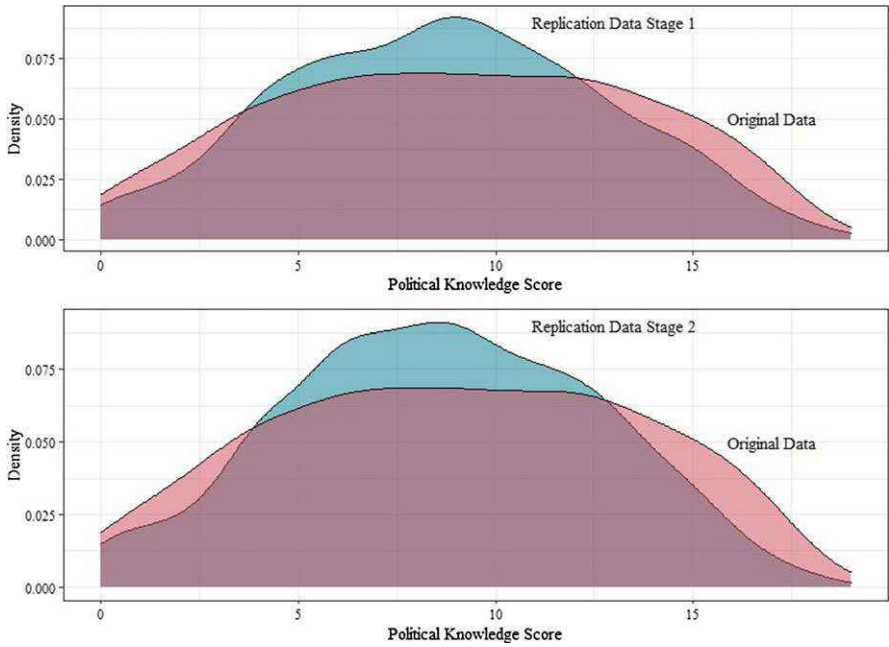
Figure 2
Frequency distribution of Political Knowledge Test scores for the replication data (blue) and the original data (pink) in Stage 1 (upper panel) and Stage 2 (bottom panel).
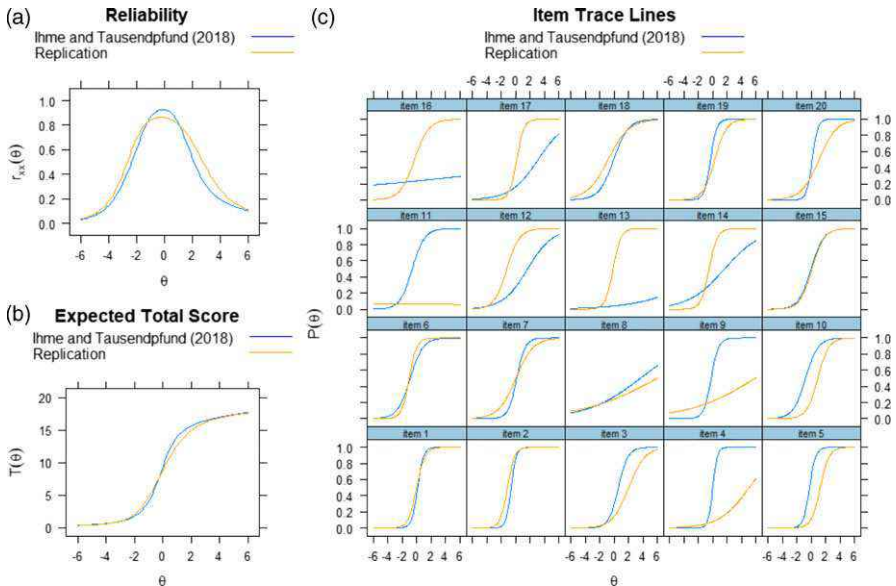

Figure 3
Results of IRT's 2PL model of the Political Knowledge Test Scores.

activation of gender stereotypes affects participants' performance in a political knowledge test. Indeed, we find evidence of absence of this effect.

We note that some elements of our study design diverged from the original study and could have contributed to the observed non-replication. Our study was conducted with American students and working adults, whereas the original study included German students. As the United States has achieved relatively lower gender parity than Germany in political empowerment (World Economic Forum 2021), one could argue that negative stereotypes about women might be more salient for Americans than Germans, undermining women's cognitive performance even in the absence of stereotype activation (e.g., in the control condition). Although we cannot rule out that some populations might be more vulnerable to gender stereotyping than others, we have reduced cultural biases as much as possible by devising a political knowledge test that was – at the same time – similar to the one used in the original study regarding the level of difficulty, as our data suggest, and relevant to the American political context. A comparison of the effect of stereotype threat on gender differences in political knowledge across countries with varying levels of gender equality would be beneficial for a better understanding of potential cultural differences in stereotype threat. Second, as a direct consequence of including working adults in our sample, it was necessary to adapt the measure of field of study to encompass the field of work. We argue, however, that this should not have contributed to the unsuccessful replication. If our measure of field of study/work would inadvertently make participants aware of their affiliation with a Politics or Non-Politics group, the effects of gender stereotype activation on performance would presumably become more salient. Instead, our results show that the field of study/work did not influence the results (Tables S16–S17). An argument can be made, however, that the extensive list of topics in our study reduced participants' self-identity with Politics. Nevertheless, adding participants' attributed importance of Politics to their study/work as a covariate in the analyses did not change results (Tables S18–S19). We have also conducted further tests restricting our sample to young and educated adults to achieve a sample more similar in composition to the respondents in the original study, but we could still not replicate the effect of stereotype activation on the gender gap in political knowledge (Table S20).

We note that our failure to replicate the effect of stereotype threat on gender differences in political knowledge is consistent with recent research efforts challenging the effect of stereotype threat on academic performance more broadly. Stoet and Geary (2012) showed that only 30% of efforts aiming to replicate the gender gap in mathematical performance do succeed. In addition, a meta-analysis investigating the effect of gender stereotype threats on the performance of schoolgirls in stereotyped subjects (e.g., science, math) indicated several signs of publication bias within this literature (Flore and Wicherts 2015). Given these results, it is plausible that the effect of gender stereotype activation might be small in magnitude and/or might be decreasing over time (Lewis and Michalak 2019).

Furthermore, we find robust evidence of a gender gap in political knowledge even after controlling for political interest. Our results validate previous accounts that the gender gap on political knowledge may be an artifact of how knowledge is conceptualized and measured and of different gender attitudes toward standard tests. In

line with previous research stating that the political knowledge gap might be artificially inflated by a disproportionate amount of men who are willing to guess rather than chose the "don't know" option – even if that might lead to an incorrect answer (Mondak and Anderson 2004) – we find that female participants attempted to answer less questions and used the "don't know" response option in the political knowledge test more frequently than their male counterparts whereas men guessed their answers more frequently than women, resulting in a larger amount of incorrect answers (Tables S8–S14). This suggests factors other than knowledge might contribute to the gender gap in political knowledge (Mondak 1999). For example, gender differences in risk taking and competitiveness (Lizotte and Sidman 2009) as well as in self-confidence (Wolak 2020) and self-efficacy (Preece 2016) may lead women to second-guess themselves and be less prone to attempt answering the questions of which they are unsure. Meanwhile, higher competitiveness and confidence in males might lead them to guess and "gain the advantage from a scoring system that does not penalize wrong answers and rewards right ones" (Kenski and Jamieson 2000, 84). Measurement non-invariance, too, appears to detrimentally affect the interpretation and validity of political knowledge scales across several sociodemographics. For example, Lizotte and Sidman (2009) and Mondak and Anderson (2004) have shown political knowledge instruments violate the equivalence assumption for gender, while Abrajano (2015) and Pietryka and MacIntosh (2013) found non-invariance across age, income, race, and education. In our own replication attempt, we also found evidence of measurement non-invariance using item response theory and showed that the magnitude of the gender systematic bias appears to be contingent on respondents' knowledge levels such that lack of equivalence by gender is stronger at average scores and weaker at the extremes of the political knowledge continuum (see Table S21 and Figure S1).

As Politics has been essentially a male-dominated field since its creation, it should not come as a surprise that current measures of political knowledge tend to favor what men typically know. Previous studies have shown that the mere inclusion of gendered items on scales of political knowledge lessens the gender gap (Barabas, Jerit, Pollock, and Rainey 2014; Dolan 2011). The investigation and validation of measures of political knowledge that capitalize on the fact that men and women might not only know different things but also may react in different ways to standard tests is paramount for a more accurate understanding of the gender gap in political knowledge and its bias.

Finally, we note that measurement issues are not unique to political knowledge and in fact are pervasive in Political Science with consequences for how we measure populism (Van Hauwaert, Schimpf, and Azevedo 2018, 2020; Wuttke, Schimpf, and Schoen 2020), operational ideology (Azevedo and Bolesta 2022; Azevedo, Jost, Rothmund, and Sterling 2019; Kalmoe 2020), and political psychological constructs such as authoritarianism, racial resentment, personality traits, and moral traditionalism (Azevedo and Jost 2021; Bromme, Rothmund, and Azevedo 2022; Pérez and Hetherington 2014; Pietryka and MacIntosh 2022). If the basic measurement properties of widely used constructs are flawed, it is likely that insights from research will be biased. Valid, invariant, and theoretically derived instruments are urgently needed for the reliable accumulation of knowledge in Political Science.

References

Abrajano, M. 2015. Reexamining the "Racial Gap" in Political Knowledge. *The Journal of Politics 77*(1): 44–54. https://doi.org/10.1086/678767

Alipourfard, N., B. Arendt, D. M. Benjamin, N. Benkler, M. M. Bishop, M. Burstein, ... **J. Wu**. 2021. Systematizing Confidence in Open Research and Evidence (SCORE). *OSF.* https://doi.org/10.31235/osf.io/46mnb

Aronow, P. M., J. Kalla, L. Orr, and J. Ternovski. 2020. Evidence of Rising Rates of Inattentiveness on Lucid in 2020. *OSF.* https://doi.org/10.31235/osf.io/8sbe4

Azevedo, F., and D. S. Bolesta. 2022. Measuring Ideology: Current Practices, Its Consequences, and Recommendations. https://measuring.ideology.flavioazevedo.com/

Azevedo, F., L. Micheli, and D. S. Bolesta. 2022. Replication Data for: Does Stereotype Threat contribute to the Political Knowledge Gender Gap? A Preregistered Replication Study of Ihme and Tausendpfund (2018). *Harvard Dataverse.* https://doi.org/10.7910/DVN/ETUUOD.

Azevedo, F., and J. T. Jost 2021. The Ideological Basis of Antiscientific Attitudes: Effects of Authoritarianism, Conservatism, Religiosity, Social Dominance, and System Justification. *Group Processes & Intergroup Relations 24*(4): 518–49. https://doi.org/10.1177/1368430221990104

Azevedo, F., J. T. Jost, T. Rothmund, and J. Sterling. 2019. Neoliberal Ideology and the Justification of Inequality in Capitalist Societies: Why Social and Economic Dimensions of Ideology are Intertwined. *Journal of Social Issues 75*(1): 49–88. https://doi.org/10.1111/josi.12310

Barabas, J., J. Jerit, W. Pollock, and C. Rainey 2014. The Question (s) of Political Knowledge. *American Political Science Review 108*(4): 840–55. https://doi.org/10.1017/S0003055414000392

Bromme, L., T. Rothmund, and F. Azevedo. 2022. Mapping Political Trust and Involvement in the Personality Space—A Meta-Analysis and New Evidence. *Journal of Personality.* https://doi.org/10.1111/jopy.12700

Burns, N., K. L. Schlozman, and S. Verba 2001. *The Private Roots of Public Action. Gender, Equality, and Political Participation*. Cambridge: Harvard University Press. https://doi.org/10.4159/9780674029088

Carpini, M. X. D., and S. Keeter. 2005. Gender and Political Knowledge. In *Gender and American Politics. Women, Men, and the Political Process*, eds. Tolleson-Rinehart S. & Josephson J. J. Armonk: Sharpe. https://doi.org/10.4324/9781315289779

Champely, S. 2020. pwr: Basic Functions for Power Analysis. *R Package Version 1.3-0*. https://CRAN.R-project.org/package=pwr

Dolan, K. 2011. Do Women and Men Know Different Things? Measuring Gender Differences in Political Knowledge. *The Journal of Politics* 73(1): 97–107. https://doi.org/10.1017/S0022381610000897

Dow, J. K. 2009. Gender Differences in Political Knowledge: Distinguishing Characteristics-Based and Returns-Based Differences. *Political Behavior* 31(1): 117–36. https://doi.org/10.1007/s11109-008-9059-8

Flore, P. C., J. Mulder, and J. M. Wicherts. 2018. The Influence of Gender Stereotype Threat on Mathematics Test Scores of Dutch High School Students: A Registered Report. *Comprehensive Results in Social Psychology* 3(2): 140–74. https://doi.org/10.1080/23743603.2018.1559647

Flore, P. C., and J. M. Wicherts. 2015. Does Stereotype Threat Influence Performance of Girls in Stereotyped Domains? A Meta-Analysis. *Journal of School Psychology* 53(1): 25–44. https://doi.org/10.1016/j.jsp.2014.10.002

Ihme, T. A., and M. Tausendpfund. 2018. Gender Differences in Political Knowledge: Bringing Situation Back In. *Journal of Experimental Political Science* 5(1): 39–55. https://doi.org/10.1017/XPS.2017.21

Kalmoe, N. P. 2020. Uses and Abuses of Ideology in Political Psychology. *Political Psychology* 41(4): 771–93. https://doi.org/10.1111/pops.12650

Kenski, K., and K. H. Jamieson 2000. The Gender Gap in Political Knowledge: Are Women Less Knowledgeable Than Men about Politics? In *Everything You Think You Know about Politics . . . And Why You're Wrong*, ed. Jamieson K. H. New York: Basic Books, 83–89.

Lenth, R. V. 2022. emmeans: Estimated Marginal Means, aka Least-Squares Means. *R Package Version 1.8.1-1*. https://CRAN.R-project.org/package=emmeans

Lewis Jr. N. A., and N. M. Michalak. 2019. Has Stereotype Threat Dissipated Over Time? A Cross-Temporal Meta-Analysis. *Psyarxiv*. https://doi.org/10.31234/osf.io/w4ta2

Lewis Jr. N. A., and D. Sekaquaptewa. 2016. Beyond Test Performance: A Broader View of Stereotype Threat. *Current Opinion in Psychology* 11: 40–3. https://doi.org/10.1016/j.copsyc.2016.05.002

Lizotte, M. K., and A. H. Sidman 2009. Explaining the Gender Gap in Political Knowledge. *Politics & Gender* 5(2), 127–51. https://doi.org/10.1017/S1743923X09000130

Ly, A., A. Etz, M. Marsman, and E. J. Wagenmakers. 2019. Replication Bayes Factors from Evidence Updating. *Behavior Research Methods* 51(6): 2498–508. https://doi.org/10.3758/s13428-018-1092-x

McGlone, M. S., J. Aronson, and D. Kobrynowicz 2006. Stereotype Threat and the Gender Gap in Political Knowledge. *Psychology of Women Quarterly* 30(4): 392–8. https://doi.org/10.1111/j.1471-6402.2006.00314.x

McGlone, M. S., and R. A. Pfiester 2007. The Generality and Consequences of Stereotype Threat. *Sociology Compass* 1(1): 174–90. https://doi.org/10.1111/j.1751-9020.2007.00021.x

Mondak, J. J. 1999. Reconsidering the Measurement of Political Knowledge. *Political Analysis* 8(1): 57–82. https://doi.org/10.1093/oxfordjournals.pan.a029805

Mondak, J. J., and M. R. Anderson 2004. The Knowledge Gap: A Reexamination of Gender-Based Differences in Political Knowledge. *The Journal of Politics* 66(2): 492–512. https://doi.org/10.1111/j.1468-2508.2004.00161.x

Ondercin, H. L., and D. Jones-White. 2011. Gender Jeopardy: What Is the Impact of Gender Differences in Political Knowledge on Political Participation? *Social Science Quarterly* 92(3): 675–94. https://doi.org/10.1111/j.1540-6237.2011.00787.x

Parsons, S., F. Azevedo, M. M. Elsherif, S. Guay, O. N. Shahim, G. H. Govaart, E. Norris, A. O'Mahony, A. J. Parker, A. Todorovic, C. R. Pennington, E. Garcia-Pelegrin, A. Lazić, O. M. Robert-son, S. L. Middleton, B. Valentini, J. McCuaig, B. J. Baker, E. Collins, . . . B. Aczel. 2021. A Community-Sourced Glossary of Open Scholarship Terms. Retrieved from https://forrt.org/glossary/

Pennington, C. R., D. Heim, A. R. Levy, and D. T. Larkin. 2016. Twenty Years of Stereotype Threat Research: A Review of Psychological Mediators. *PLoS ONE* 11(1): e0146487. https://doi.org/10.1371/journal.pone.0146487

Pennington, C. R., D. Litchfield, N. McLatchie, and D. Heim 2018. Stereotype Threat May Not Impact Women's Inhibitory Control or Mathematical Performance: Providing Support for the Null Hypothesis. *European Journal of Social Psychology* 49(4): 717–34. https://doi.org/10.1002/ejsp.2540

Pérez, E., and M. Hetherington. 2014. Authoritarianism in Black and White: Testing the Cross-Racial Validity of the Child Rearing Scale. *Political Analysis* 22(3): 398–412. https://doi.org/10.1093/pan/mpu002

Pietryka, M. T., and R. C. MacIntosh. 2013. An Analysis of ANES Items and Their Use in the Construction of Political Knowledge Scales. *Political Analysis* 21(4): 407–29. https://doi.org/10.1093/pan/mpt009

Pietryka, M. T., and R. C. MacIntosh. 2022. ANES Scales Often Do Not Measure What You Think They Measure. *The Journal of Politics* 84(2): 1074–90. https://doi.org/10.1086/715251

Preece, J. R. 2016. Mind the Gender Gap: An Experiment on the Influence of Self-Efficacy on Political Interest. *Politics & Gender* 12(1): 198–217. https://doi.org/10.1017/S1743923X15000628

Pruysers, S., and J. Blais. 2014. Anything Women Can Do Men Can Do Better: An Experiment Examining the Effects of Stereotype Threat on Political Knowledge and Efficacy. *The Social Science Journal* 51(3): 341–9. https://doi.org/10.1016/j.soscij.2014.05.005

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Schmader, T., M. Johns, and C. Forbes. 2008. An Integrated Process Model of Stereotype Threat Effects on Performance. *Psychological Review* 115(2): 336–56. https://doi.org/10.1037/0033-295X.115.2.336

Stoet, G., and D. C. Geary 2012. Can Stereotype Threat Explain the Gender Gap in Mathematics Performance and Achievement? *Review of General Psychology* 16(1): 93–102. https://doi.org/10.1037/a0026617

Van Hauwaert, S. M., C. H. Schimpf, and F. Azevedo. 2018. Public Opinion Surveys: Evaluating Existing Measures. In *The Ideational Approach to Populism*: Routledge, 154–172. https://doi.org/10.4324/9781315196923-7

Van Hauwaert, S. M., C. H. Schimpf, and F. Azevedo. 2020. The Measurement of Populist Attitudes: Testing Cross-National Scales Using Item Response Theory. *Politics* 40(1): 3–21. https://doi.org/10.1177/0263395719859306

Verhagen, J., and E. J. Wagenmakers. 2014. Bayesian Tests to Quantify the Result of a Replication Attempt. *Journal of Experimental Psychology: General* 143(4): 1457–75. https://doi.org/10.1037/a0036731

Walton, G. M., and G. L. Cohen 2003. Stereotype Lift. *Journal of Experimental Social Psychology* 39(5): 456–67. https://doi.org/10.1016/S0022-1031(03)00019-2

Wolak, J. 2020. Self-Confidence and Gender Gaps in Political Interest, Attention and Efficacy. *The Journal of Politics* 82(4): 1490–501. https://doi.org/10.1086/708644

World Economic Forum. 2021. The Global Gender Gap Report 2021. Retrieved from https://www.weforum.org/reports/ab6795a1-960c-42b2-b3d5-587eccda6023

Wuttke, A., C. Schimpf, and H. Schoen. 2020. When the Whole Is Greater Than the Sum of Its Parts: On the Conceptualization and Measurement of Populist Attitudes and Other Multidimensional Constructs. *American Political Science Review* 114(2): 356–74. https://doi.org/10.1017/S0003055419000807

Zwaan, R. A., A. Etz, R. E. Lucas, and M. B. Donnellan 2018. Making Replication Mainstream. *Behavioral and Brain Sciences* 41: E120. https://doi.org/10.1017/S0140525X17001972