American Political Science Review (2026) 120, 4, 1561–1569
doi:10.1017/S0003055426101828 © The Author(s), 2026. Published by Cambridge University Press on behalf of American Political Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Replication/Reappraisal

Still Instrumentally Inclusive

STUART J. TURNBULL-DUGARTE *University of Southampton, United Kingdom*
ALBERTO LÓPEZ ORTEGA *Vrije Universiteit Amsterdam, Netherlands*

Do individuals in Western democracies shift their views on LGBT+ inclusion when exposed to opposition from Muslim out-groups? And is this "homonationalist" responsiveness stronger among anti-immigration individuals? Our article "Instrumentally inclusive […]" addressed these questions using representative data from Britain (Study 1) and crowd-sourced data from Spain (Study 2). We revisit Study 2 in response to concerns about post-stratification weighting a convenience sample and erroneously inconsistent use of robust standard errors. Varying population reference parameters, weight caps, variance estimators, subgroup definitions, and covariate adjustments, we find strong support for the original findings. Neither reasonable alternative weighting schemes nor corrected variance estimators substantively change the results. In fact, the originally published point estimates are equal to or smaller than the median across alternative specifications. Consistent with the Article, we find that treatment increases support for LGBT+ inclusion in Spain, without consistent identifiable variation by immigration attitudes.

In our original paper "*Instrumentally* Inclusive: The Political Psychology of Homonationalism" (hereafter, the *Article*) we presented evidence from two survey experiments (Turnbull-Dugarte and Ortega 2024). Both experiments—the primary experiment using a representative sample in Britain (Study 1) and a replication experiment using a crowd-sourced (convenience) sample in Spain (Study 2)—randomly presented individuals with vignettes highlighting opposition to LGBT+ inclusive education in schools among Muslims or nondescript individuals. The *Article* presents evidence that exposure to vignettes featuring Muslims resulted in an increase in support for LGBT+ inclusive education.

In Study 1, the *Article* identified a statistically significant difference in effects between pro- and anti-immigration respondents based on a dichotomous stratification of the pretreatment moderator. Anti-immigration respondents liberalized their views on LGBT+ education in response to treatment whereas pro-immigration respondents did not (*Article*, Figure 4). An alternative stratification of the moderator based on trichotomous groups—low (0–3), middle (4–6), and high (7–10) values—was included to demonstrate the results across distinct means of stratifying respondents. The results of this alternative grouping reported similar findings (*Article*, Figure A11).

In the replication Study 2, the *Article* found that respondents liberalized their views on LGBT+ education in response to treatment on average, but did *not* identify statistically significant differences between anti- and pro-immigration respondents based on a dichotomous split of the sample. In other words, the liberalization in views in response to treatment was *equal* among both groups when split down the middle (*Article*, Figure 7). The alternative specification of the moderator based on a trichotomous grouping did, however, find that the effect of the treatment was significantly greater among the most anti-immigration (lowest tercile) compared to the most pro-immigration (highest tercile). The effect of treatment was significantly distinct from zero and positive among both groups while there was no identifiable effect among those with pretreatment immigration attitudes around the middle (*Article*, Figure A13).

De Kadt (the *Commentator*) has produced a contribution (hereafter, the *Comment*) to critique some of the empirical components of the *Article*, specifically in relation to replication Study 2 (De Kadt 2025). We appreciate the engagement with our work and for the effort involved in reanalyzing our data. As we indicated to the *Commentator* in our initial correspondence[1]

Corresponding author: Stuart J. Turnbull-Dugarte, Associate Professor, Department of Politics and International Relations, University of Southampton, United Kingdom, s.turnbull-dugarte@soton.ac.uk.
Alberto López Ortega, Assistant Professor, Faculty of Social Sciences and Humanities, Vrije Universiteit Amsterdam, Netherlands, a.lopez.ortega@vu.nl.

Handling editor: Sebastian Karcher.

Received: September 26, 2025 UTC; revised: May 11, 2026 UTC; accepted: June 25, 2026 UTC. First published online: July 28, 2026 UTC.

[1] The *Commentator* shared an email with us on October 2, 2024 signaling some concerns with our article. We responded on October 9 (see Supplementary Material Correspondence I) signaling our desire to alleviate some of their concerns and to form a collaborative path forward including pursuing a joint replication using new data. We did not receive any further communication in response. After we did not hear from the *Commentator* or the *APSR*, we later contacted the *APSR* editorial team ourselves soliciting guidance on how to submit a corrigendum based on the two identified coding errors in the

https://doi.org/10.1017/S0003055426101828 Published online by Cambridge University Press

when they communicated their concerns with us, we believe reproductions and replications are essential components of knowledge production, particularly when they serve to correct priors and update the state of shared knowledge within the discipline. This was precisely the motivation behind including a replication (Study 2) in the original *Article.* Our primary objective remains understanding whether or not the claims related to Study 2 in the *Article* are "right": do individuals in Spain update their views on LGBT+ inclusion in response to out-group cues and is this homonationalist responsiveness greater among anti-immigration individuals? Ultimately, we want to know if our current understanding is correct or not, even if that means correcting potentially flawed priors.

We wish to explicitly recognize the two errors in the *Article* identified in the *Comment.* First, there are two instances (Figures 3 and 6 in the *Article*) where illustrations based on logistic regression models were erroneously labeled as being sourced from OLS regression models. Second, the application of robust standard errors was mistakenly not uniform. We regret these errors in our code and thank the *Comment* for bringing these errors to our attention. Despite these errors, the *Comment* and the reanalysis of the *Article* that it reports do not change the empirical claims presented in the *Article*, nor do they alter the conclusions or the theoretical insights supported by the empirical evidence. The *Comment* mischaracterizes our analytical choices as "ad hoc," relies on assumptions about the value of unadjusted convenience samples that do not align with common practice, and overstates the implications on the *Article*'s conclusions. In what follows, we demonstrate that the claims presented in the *Article* are robust across a diverse universe of alternatives—including researcher degrees of freedom related to weighting choices—and are *not* influenced by the two errors that the *Comment* correctly identifies. Specifically, we reanalyze the empirical claims from Study 2 via the estimation of 2,970 alternative approaches to analyzing the data and testing our hypotheses. Across these different specifications, we find a consistent pattern. Individuals become more supportive of LGBT+ education when randomly exposed to treatment and, in contrast to Study 1 but *as reported in the Article's description of Study 2*, we find no consistent statistically significant differences, *on average,* based on pretreatment immigration attitudes.

PRIMARY CONCERN: USE OF POST-STRATIFICATION WEIGHTS IN A DEMOGRAPHICALLY UNREPRESENTATIVE CONVENIENCE SAMPLE

The weighting dispute in the *Comment* rests on a prior question: what is the appropriate estimand for Study 2? Our theory concerns the political psychology of adults in Spain and, given our expectation of subgroup variation, differences between those with more positive or negative views toward immigration. The target of inference is therefore a *population average treatment effect* (PATE), not a sample average treatment effect (SATE) that describes only those who happened to take part in our survey. Miratrix et al. (2018) are explicit that when the goal is PATE estimation, post-stratification weighting of a convenience sample is not merely permissible but methodologically required; unweighted estimates are only equivalent to weighted ones when the sample is already broadly representative of population parameters—a condition that, as Figure 1 makes plain, the sample for Study 2 decisively fails to meet. The *Comment*'s preferred unweighted analysis recovers a SATE for a sample dominated by young, urban respondents. That is a legitimate quantity, but it is not the one our theory predicts nor the one our *Article* claims to estimate.

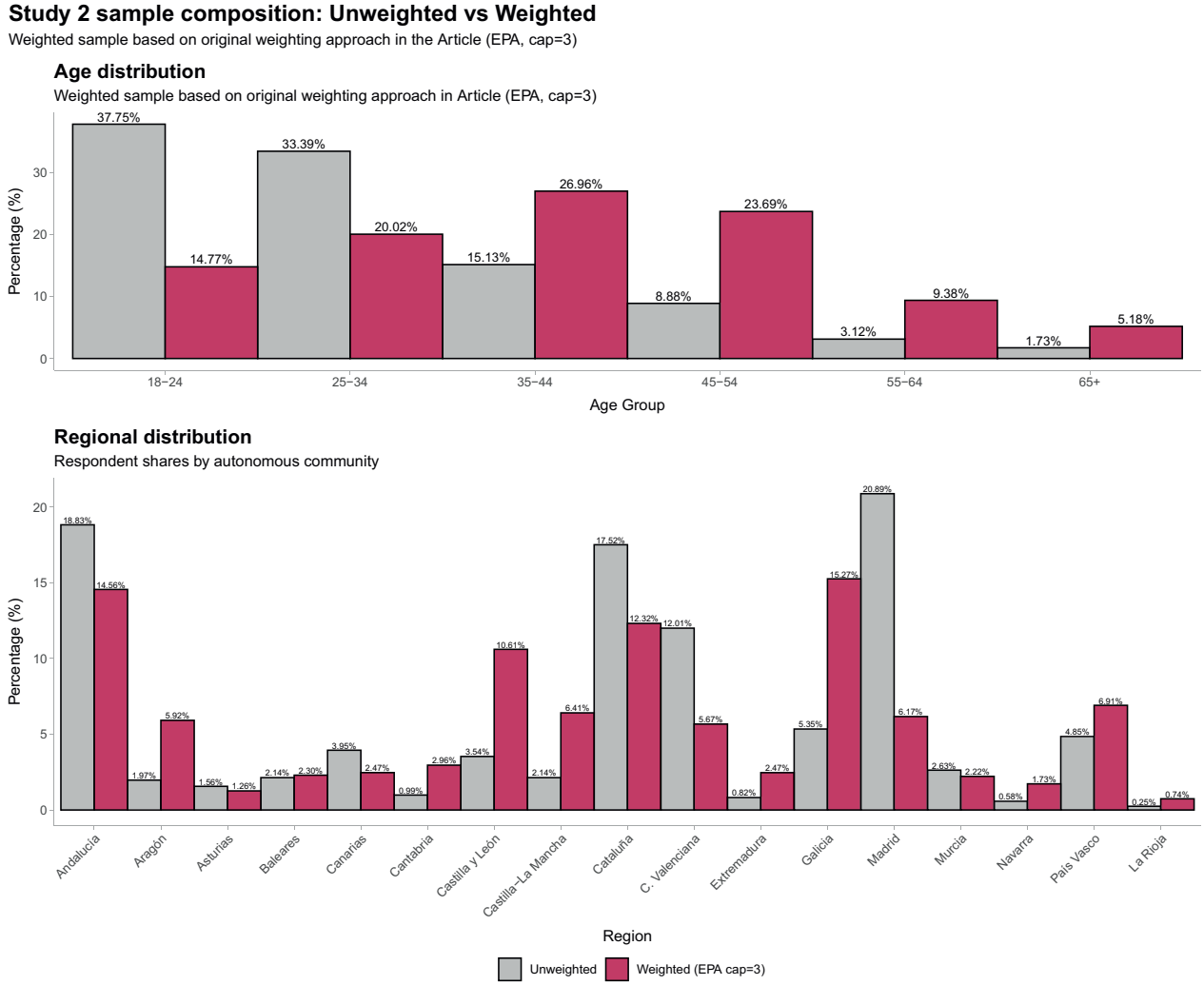
The *Comment* suggests that applying post-stratification weights to a convenience sample is, in itself, problematic and generally discouraged. This claim is not supported by a consensus in the methodological literature. Franco et al. (2017) recommend reporting both weighted and unweighted estimates when possible, but at no point do they argue against weighting convenience samples as the *Comment* claims. As Franco et al. (2017) highlight, it is not uncommon in political science research to apply post-stratification weights to demographically skewed convenience samples where the aim is to approximate population parameters. A quick search among the replication files of recent publications in high-ranking political science journals reveals that post-stratification weighting is commonplace. See, for example, Baisley and Albaugh (2025), Barari, Lucas, and Munger (2025), Broockman and Kalla (2025), Culpepper, Jung, and Lee (2024), Grose, Malhotra, and Van Houweling (2015), Hainmueller and Hopkins (2015), Marsh and Peterson (2026), Matanock and García-Sánchez (2018), Mikkelborg (2025), Schaeffer, Krakowski, and Olsen (2024), and Wittenberg et al. (2023) for a non-exhaustive list.

The *Comment* describes the decision to use weights in Study 2 as "inconsistent with the analysis of Study 1, and this inconsistency is not explained or justified." This is a mischaracterization: the *Article* clearly states that Study 1 data are based on a representative sample (*Article*, 1366) and that Study 2 is based on a crowdsourced sample that we weight to approximate the population (*Article*, 1370). This difference reflects different sampling designs that are explicitly detailed in the *Article*, not an ad hoc choice. Had we been able to finance the recruitment of a representative sample for our replication Study 2, this would have been our preferred approach to replicating Study 1. As that was not the case, we corrected the sub-optimal sample with weights, a standard methodological approach in political science survey experiments (Broockman and Kalla 2025; Hainmueller and Hopkins 2015; Mikkelborg 2025), and explicitly stated doing so in the *Article.* Given resource constraints, we relied on a crowdsourced sample for Study 2 sourced from Prolific. While Prolific can provide representative samples for the United Kingdom and the United States, it does not,

Article (see Supplementary Material Correspondence II). It was only after our initial correspondence, that we became aware that a *Comment* on the *Article* was forthcoming and we were invited to share a response within a 60-day deadline.

https://doi.org/10.1017/S0003055426101828 Published online by Cambridge University Press

FIGURE 1. Demographic Skew in Unweighted Sample and Correction with Weights



Note: Comparison of age and regional distribution in the unweighted crowd-sourced sample from Prolific against the Spanish adult population (EPA 2022 targets) and the post-stratification weighted sample. Full distribution statistics are reported in Table A1 in the Supplementary Material.

currently, do so for other countries. The sample of respondents in Study 2 was severely skewed toward younger individuals and those based in Spain's capital Madrid. As openly stated in the *Article* (1370), we applied post-stratification weights to correct for the demographic skew of the data.

The *Comment* characterises the bimodal shape of our weight distribution as "peculiar." In fact, this pattern is a predictable and direct consequence of applying standard weighting procedures to a sample where large subgroups in the raw sample are far from population targets (e.g., under-representation of older respondents and under-representations of regions outside of Madrid) and when weights are capped. In this instance, a bimodal weighting distribution naturally occurs: those over-represented will receive essentially no weight whereas those severely underrepresented will receive the maximum weight. Despite the *Comment*'s pointed insinuations, and the description of the weight distribution used as "highly unusual," the *Comment* made a clear decision not to (i) demonstrate what a usual weighting distribution should look like given the sample and how those used in the *Article* deviates from it and (ii) create original weights from different population parameters in Spain and compare them to those used in the *Article.* As we demonstrate in what follows, any simple comparison of the weights used in the *Article* to a diverse slate of alternative weighting approaches results in comparable results and consistent inferences to those original reported.

Sample Weighting Procedure

Since the *Article* did not provide sufficient detail on how the weights were constructed, we remedy this here. To enhance the representativeness of the convenience

https://doi.org/10.1017/S0003055426101828 Published online by Cambridge University Press

sample from the replication Study 2, we implemented a post-stratification weighting procedure using iterative proportional fitting (raking). As our population benchmark, we used the Spanish Labor Force Survey (*Encuesta de Población Activa*, EPA) from the first quarter of 2022, which provides nationally representative estimates of the Spanish adult population. We adjusted our sample to match the EPA population parameters on three demographic dimensions: gender (male/female), age (18–24, 25–34, 35–44, 45–54, 55–64, and 65+), and region (Spain's Autonomous communities, excluding Ceuta and Melilla). The raking algorithm, implemented with the anesrake package in R (Pasek 2018), iteratively adjusted weights to simultaneously match all target distributions, capping individual weights at 3 to avoid excessive influence from any single observation. The cap at 3 is the default option in anesrake.

Figure 1 reveals why post-stratification weighting was essential for our analysis. Our unweighted crowd-sourced sample diverged dramatically from the Spanish adult population across key demographic dimensions (see Table A1 in the Supplementary Material) but particularly age. Most strikingly, individuals aged 18–34 were dramatically over-represented and comprised 71.1% of our unweighted sample. Conversely, those aged 65 and older represented only 1.7% of respondents. Regional imbalances were equally pronounced (although with variation depending on the population parameter, see Table A1 in the Supplementary Material). As shown in Figure 1, post-stratification successfully reduced, without fully correcting, these demographic imbalances. While the weight cap of 3 prevented perfect convergence to population parameters, particularly for the most extreme age imbalances, the weighted sample represents a substantial improvement in demographic representativeness.

Testing the Stability of Results across Alternative Population Weighting Schemes

Given potential concerns about the sensitivity of our findings to population weighting decisions, we conducted a comprehensive robustness analysis using three different Spanish population benchmarks and six different weight capping values. As weighting decisions may represent an additional researcher degree of freedom (Franco et al. 2017), our goal here is to verify whether the treatment effects reported in the *Article* remain stable across multiple alternative weighting approaches and if those reported differ systematically from those that might be observed via the research forking path. They do not: the point estimates reported in the *Article* are smaller than the median point estimate produced by alternative specifications.

We constructed three separate weighting schemes drawing from authoritative Spanish sources. The first uses the Spanish Labour Force Survey (*Encuesta de Población Activa*, EPA) from the first quarter of 2022, which provides quarterly employment and demographic data for the Spanish population. This is the weighting procedure applied in the *Article.* The second employs official population figures from the National Statistics Institute (*Instituto Nacional de Estadística*, INE) as of July 2022, representing the most authoritative census-based population estimates. The third relies on the Center for Sociological Research (*Centro de Investigaciones Sociológicas*, CIS) Barometer Survey No. 3366 from June 2022, a major nationally representative political attitudes survey. Each weighting scheme adjusts our sample to match population distributions across gender, six age categories, and Spain's 17 autonomous communities. Beyond varying population benchmarks, we also tested the sensitivity of results to the weight truncation decision. While our primary analysis employed the default cap of 3 (following standard practice in anesrake), we varied this constraint, testing caps of 2 (most restrictive), 3 (default), and 4, 5, 10, and 100 (effectively uncapped). This yields 18 distinct weight-specification combinations (3 benchmarks × 6 caps), allowing us to assess whether our findings depend on particular weighting choices.

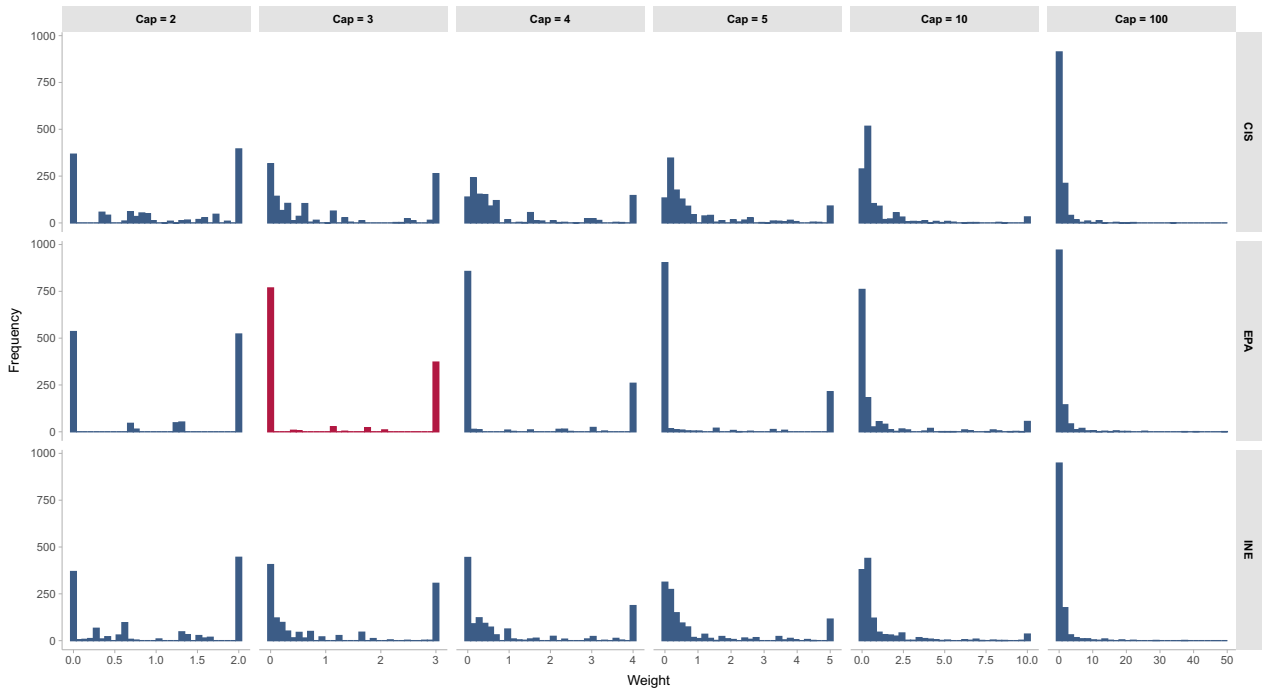
Figure 2 illustrates how weight distributions evolve across these alternative population benchmark sources and weighting caps. The bimodal pattern (with concentrations near 0 and at the cap value, 3) reveals the fundamental challenge of our crowd-sourced sample. Young respondents, who dominate our sample, receive minimal weights (often below 0.01) to down-weight their over representation, while the few older respondents receive maximum weights at the cap threshold. This pattern persists across all three population benchmarks, confirming that the weight distribution is driven by the underlying sample composition rather than the choice of the demographic reference target. As the cap increases from 2 to 100, the distributions become progressively more dispersed. With cap = 2, weights are tightly constrained between 0 and 2, creating relatively uniform influence across observations but limiting demographic correction. With cap = 100, the EPA and INE distributions appear nearly identical, while CIS weights show slightly different patterns, particularly in the right tail, reflecting its distinct population profile.

In Figure 3, we demonstrate that treatment effects are stable across weighting alternatives (see also Appendices A.1 and A.3 of the Supplementary Material). Figure 3 reports treatment effects, with robust (HC3) standard errors, that vary the 18 different weighting approaches alongside diverse means of covariate adjustment.[2] The *Comment*'s primary concern is that the conclusions depend on "ad hoc" weighting choices. We empirically refute this: any sensible approach to demographic adjustment yields consistent conclusions about the instrumental nature of LGBT+ support in response to out-group cues. Our weighting procedure, regardless of its specific implementation, corrects this bias and approximates the treatment effect we could anticipate in a demographically balanced sample. It is also worth highlighting that the distributional position of the *Article*'s estimates are inconsistent

[2] Multiverse analysis also varying different variance estimators—classic standard errors, HC3 standard errors, and survey-robust standard errors are reported in Appendix A.3 of the Supplementary Material. This, more comprehensive multiverse analysis, renders a total of 8,910 specifications.

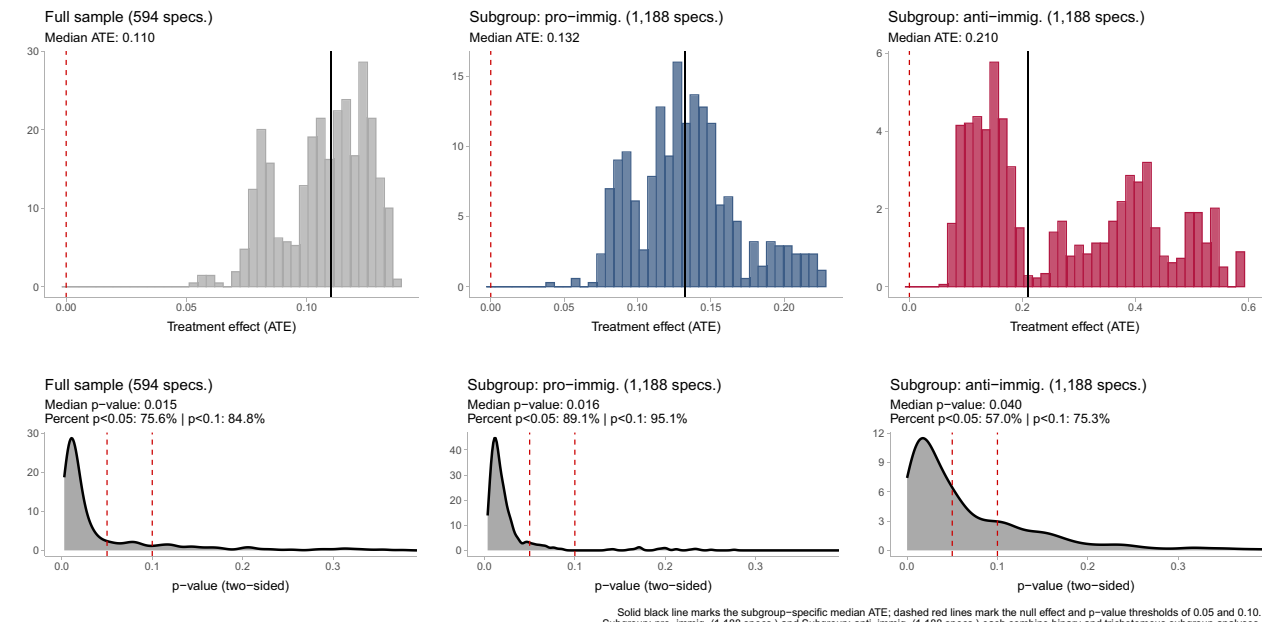
https://doi.org/10.1017/S0003055426101828 Published online by Cambridge University Press

FIGURE 2. Distribution of Survey Weights by Reference Scheme and Cap Value



Note: Each panel displays the distribution of survey weights for one of 18 weighting specifications—three population benchmarks (EPA, INE, and CIS) crossed with six weight caps (2, 3, 4, 5, 10, and 100). The bimodal pattern reflects the severe under-representation of certain respondents in the crowd-sourced sample. The highlight panel, EPA at Cap=3 indicates that used in the *Article.*

FIGURE 3. Treatment Effects and *p*-Values from 2,970 Alternative Means of Testing the Hypotheses



Note: Summary of 2,970 specifications testing the treatment effect in Study 2 across the full sample and subgroups stratified by pretreatment immigration attitudes (both dichotomous and trichotomous), varying weights applied and covariate adjustment. Each point represents one specification.

https://doi.org/10.1017/S0003055426101828 Published online by Cambridge University Press

with *Comment*'s charge of selective analysis. Cherry-picking, the deliberate choice of specifications that favor one's preferred conclusion, would predict that the reported estimates sit at the upper tail of the distribution of plausible alternatives. The opposite is true. As Figure 3 shows, the originally reported point estimates belong to the *lower* end of the distribution of estimates generated by the 2,970 alternative specifications. The *Article*'s estimates are, if anything, conservative relative to the space of defensible analytical choices.

SECONDARY CONCERNS

In addition to the issue related to the use of weights in a demographically skewed crowd-sourced sample, the *Comment* raises a number of adjacent concerns. We interpret these concerns as threefold. First, the *Comment* identifies an incorrect description of the estimations visualized in Figures 3 and 6 in the *Article.* These figures are erroneously described as being based on an ordinary least squares (OLS) regression. While the corresponding tables referenced within each figure (*Article*, Tables A7 and A9, respectively) do indeed rely on OLS, the visualizations themselves are based on logistic regression models. This was a coding mistake that likely emerged during revision processes and updating analyses to reflect OLS-versus-logistic regression discussions with reviewers and we regret not updating the figure accordingly to match the tables. Whilst this is obviously an error worth correcting that should not have happened, we also view it as irrelevant to changing the empirical or theoretical conclusions: the error has no bearing on the results and our in-text discussion of the point estimates in the *Article* reflects those reported in the OLS models.

Second, the *Comment* identifies that some of the reported coefficients rely on classical standard errors and not robust standard errors. This was a coding error that we did not recognize in the audit of our code.[3] The *Comment* correctly identifies some instances where robust standard errors were erroneously *not* called. This was a mistake and one that could have been avoided had we relied on the estimatr package (Blair et al. 2025) and the corresponding lm_robust function as standard.

We disagree with the *Comment*, however, when it comes to whether this mistake influences the findings. Our original interpretation of the results from Study 2 in the *Article* was that treatment induced a positive effect on the outcome (scaled 0–1) among both pro-immigration respondents (CATE $= 0.112$) and anti-immigration respondents (CATE $= 0.102$) but there was *no* significant difference *between* the two when the sample was stratified dichotomously. In other words, the effect of treatment was statistically symmetrical for both subgroups (Gelman and Stern 2006).[4] Based on a trichotomous split, we observed significantly larger effects for the most anti-immigration respondents (CATE $= 0.32$) compared to the most pro-immigration respondents (CATE $= 0.11$). The *Comment*'s claim that "in none of the alternative specifications is the interaction term between treatment and anti-immigration sentiment statistically significant, despite this being the authors' key theoretical prediction" is flawed. Not only does it mischaracterize our own conclusion of Study 2 as showing no consistent evidence of effect homogeneity (i.e., we transparently stated that the evidence went *against* our expectations in the original *Article*), it also actively ignores instances where estimations did observe significant heterogeneous effects. Our interpretation remains the same regardless of standard error choice as we demonstrate in Figure 4.

Correcting for robust standard errors (the appropriate variance estimator) does not influence the magnitude of our identified effects for each group, nor the difference between the two, across alternative means of stratification. The *Comment* interprets the reduced significance for the anti-immigration group using the dichotomous stratification as indicative of evidence against our theory. We do not, however, interpret the point estimate of 0.102 for the dichotomous anti-immigration group (representing a 21% increase relative to the control group baseline) on a 0–1 scale as having *no* effect. *P*-values on either side of the selected significance threshold allow us to make a claim about whether an effect is statistically distinguishable from zero, but do not indicate that the effect itself is null (Altman and Bland 1995; Rainey 2014). The 0.102 point estimate remains statistically symmetrical to that observed among the full sample, and the dichotomous pro-immigration subgroup. In other words, the effect of treatment for the dichotomous anti-immigration subgroup is qualitatively and statistically symmetrical to that of the dichotomous pro-immigration subgroup even if the former may be, under some specifications, indistinguishable from zero (Gelman and Stern 2006). The point-estimate from all three models—full sample, anti-immigration, and pro-immigration—are also of a sizeable magnitude that is very unlikely to be observed under the null and the effect among the most anti-immigration tercile subgroup is even stronger at 0.32 ($p < 0.001$).

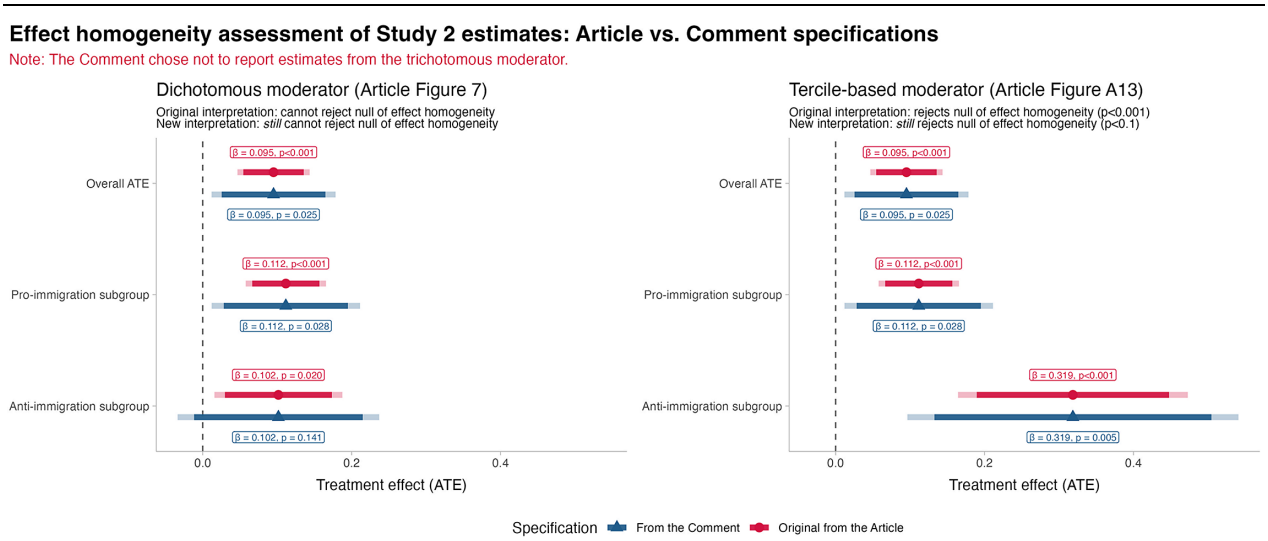
In Figure 5, we empirically demonstrate that the observed point estimate for the dichotomized anti-immigration subgroup, rather than being an effect of

[3] Our analysis and reporting relied on the robust=TRUE call. This call was correct for the jtools package (Long 2024) used for the data visualizations, but incorrect for the modelsummary package (Arel-Bundock 2022) used to create tables. For example—summary (model, robust=TRUE)—as executed via jtools correctly reports robust standard errors. Correct execution via modelsummary would be modelsummary(model, vcov = "robust"). Our code, which ran without errors, conflated the two packages and was misspecified in the following way: modelsummary(model, robust = TRUE). Replication code is available at Turnbull-Dugarte and López Ortega (2026).

[4] "Treated respondents in Spain, **regardless of whether or not they held nativist views**, reported support for LGBT+ education that was 10 or 11 percentage-points higher, respectively, than their peers with similar (non)-nativist views" (*Article*, 1372, emphasis not in original).

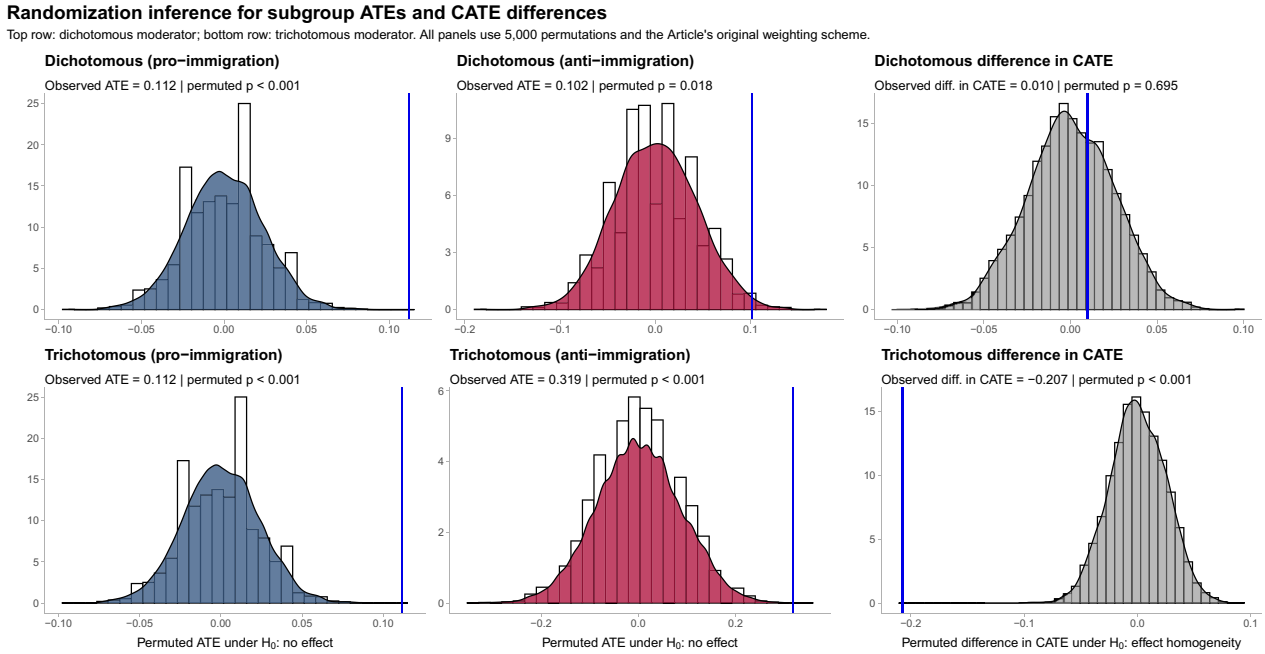
https://doi.org/10.1017/S0003055426101828 Published online by Cambridge University Press

FIGURE 4. Comparing *Article* and *Comment* Estimates



Note: Point estimates with 95% and 90% confidence intervals. Original weighting scheme applied. Article: Classical OLS SE. Comment: HC3 robust SE. Null hypothesis of effect homogeneity: CATE (Anti) = CATE (Pro).

FIGURE 5. Comparing the Identified CATE Distribution under (i) the Null of No Effect and (ii) the Null of Effect of Homogeneity



Note: Recall that the Comment chose not to report the results of the tercile-based heterogeneity tests. ATE panels show permutation distributions for subgroup treatment effects. Vertical reference line indicates identified estimate. Table output in Appendix B1 of the Supplementary Material.

zero, is a statistically improbable effect under the null of no effect. Using the ritest package (McDermott 2022), we conducted randomization inference to assess the probability of obtaining the observed point estimates under the null of no treatment effect. For each model (full sample and each subgroup in Study 2), we generated 5,000 permutations of the treatment assignment and re-estimated the effect to form an empirical null

https://doi.org/10.1017/S0003055426101828 Published online by Cambridge University Press

distribution. The resulting permutation *p*-values indicate that the observed effects are statistically rare under the null. For the anti-immigration subgroup (the subgroup analysis where the *Comment* is concerned with a null effect), the permutation *p*-value was 0.018, meaning that in fewer than 1.8% of shuffled datasets did the effect size meet or exceed the observed value in absolute terms. The permuted 95% confidence interval for this *p*-value ranged from 0.0145 to 0.022, reinforcing that the result remains unlikely to arise from chance even when accounting for simulation uncertainty.[5]

As shown in Appendix B of the Supplementary Material, where we consider the trichotomous operationalization of the moderator reported in the *Article* (Figure A13) but absent in the *Comment*'s reanalysis, permutations reveal that the probability of observing the identified effect under the null is less than 1 in 5,000 (permutation *p*-value of 0.0002). The *Comment*'s empirical contribution is that, under some specifications, the identified point estimate is not distinguishable from zero at conventional thresholds. We would highlight, however, that the robustness and sensitivity analyses reported in the *Article*'s original supplementary material, including multiverse specification curves, also transparently indicated instances where this was the case (see *Article*, Appendix G).[6] The claim made in the *Comment* that the inferences made are sensitive to ad hoc selection of standard errors is not supported by the data.

Third and finally, the *Comment* expresses some taste-based concerns related to the data visualizations. Specifically, the *Comment* does not agree with the choice not to report confidence intervals for the predicted outcomes even when the average marginal effects are reported with confidence intervals; and the *Comment* disagrees with the overlaid jittered plot of predicted outcomes and instead would have preferred to observe the overlaid underlying data. These are reasonable stylistic preferences but they are, ultimately, preferences rather than errors and at no point do our visualizations incorrectly label what is displayed beyond the logistic-versus-OLS error already acknowledged. In all figures, estimand uncertainty is reported, and readers are directly pointed toward the corresponding table where standard errors, and *p*-values are reported. The in-text description of point estimates also explicitly refers to statistical significance where appropriate.

CONCLUSION

Despite the *Comment*'s criticism and select language, we remain confident in the inferences from replication Study 2 in the *Article.* We find support for our claims about the effect of out-group cues on support for LGBT+ inclusion across the full sample of Study 2 as well as within the samples stratified based on pretreatment attitudes toward immigration. This is true whether we stratify respondents dichotomously or based on a comparison between those with the highest and lowest values of the moderator. In the *Article*, we observed significant variation between pro- and anti-immigration respondents in Study 1, and we did *not* observe consistently significant variation between these groups of respondents in the replication Study 2. The *Comment*'s reproduction has no impact on the claims produced by the *Article.* Moreover, the coefficients reported in the *Article* belong to the lower-to-median end of the distribution of identified effects and, as such, the *Comment*'s claim that the analysis reported in the *Article* selectively presented specifications that supported the author's theory is not supported. As we communicated to the *Commentator* in our correspondence (see Supplementary Material Correspondence I), replications and reproductions provide significant value added to science, particularly when they correct erroneous priors and help us gain a better understanding of the "correct" answer. The *Comment* points out where helpful clarification of the survey weights could have been added to the article but it does not update our priors.

Should instrumental inclusion among the population and, in particular, among nativists (the theorized *estimands* of interest) exist as we argue it does, then we would, of course, expect to identify similar *estimates* of these estimands in other settings. In this regard, the estimates we report can be interpreted as individual *observations.* As a result of the *Article*'s initial contribution, subsequent replications and extensions by others, using original data, provide additional observations of this estimand that are congruent with the original estimates reported in the *Article.* Raffaelli (2025), for example, presents individuals in Britain with fictitious social media posts which include part of the same visual stimuli in the *Article* and finds, among other novel additions, that respondents update their support for LGBT+ inclusion when exposed to anti-LGBT+ content associated with Islam. Mainz (2026), on the other hand, uses data from the United States to demonstrate that a similar experimental setup produces empirical support for instrumental inclusion in the case of inimical threats to women rather than sexual minorities. We have also applied the same theoretical test to a preregistered conjoint set-up in the Netherlands and Germany, where not only do we manipulate whether LGBT+ positions are associated with Islam, but we also randomize the direction (support vs. oppose) of these positions, among other characteristics (López Ortega and Turnbull-Dugarte 2025). Congruent with our theory, individuals update their views on LGBT+ rights *instrumentally* and regardless of the direction, to disidentify from the ethnic out-group. This rejection is also significantly higher among anti-immigration respondents. So far, this combined body of work has

[5] Because the permutation distribution is estimated via simulation, *p*-values and associated confidence intervals will vary slightly across runs due to random resampling variability. These differences are expected and reflect the inherent simulation error of the finite number of permutations, not any substantive change in the underlying result.

[6] The supplementary material for the *Article* already included several specification curves (see *Article*, Appendix G) which demonstrated the sensitivity of the results from the experiment to different forms of covariate adjustment and alternative means of stratifying respondents based on their pretreatment immigration attitudes. These sensitivity tests transparently identified and reported specifications where the identified treatment effects were statistically indistinguishable from zero.

https://doi.org/10.1017/S0003055426101828 Published online by Cambridge University Press

identified effects in Britain, Germany, the Netherlands, Spain, and the USA. As we signaled to the *Commentator* in our invitation to engage in new data collection, we encourage researchers to pursue further replications and extensions in similar, and indeed more diverse contexts, to understand under what conditions the dynamics of instrumental inclusion are transferable. We encourage this enthusiastically, emphatically, and consistently as the systematic understanding of an estimand can only be reinforced when repeated empirical estimates of that estimand are produced. We fundamentally disagree with the *Comment*'s premise that the findings of our replication Study 2 "evaporate" under reasonable alternative specifications. Our interpretation is that the *Comment*'s analysis leaves the core findings largely unchanged and the notion that the results from Study 2 disappear is not reflected in the evidence. As for whether it has corrected our theoretical understanding or expectations regarding the instrumental inclusion of citizens and of nativists in particular, it has not.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0003055426101828.

DATA AVAILABILITY STATEMENT

Research documentation and data that support the findings of this study are openly available at the American Political Science Review Dataverse: https://doi.org/10.7910/DVN/EWMIKZ.

ACKNOWLEDGEMENTS

The authors are grateful for the feedback from the many anonymous reviewers as well as that of the handling editor.

ETHICAL STANDARDS

The authors affirm this research did not involve human participants.

REFERENCES

Altman, Douglas G., and J. Martin Bland. 1995. "Statistics Notes: Absence of Evidence Is Not Evidence of Absence." *British Medical Journal* 311 (7003): 485.
Arel-Bundock, Vincent. 2022. "Modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23.
Baisley, Elizabeth, and Quinn M. Albaugh. 2025. "Gender Cues, Attributions and Stereotyping of Transgender and Nonbinary Politicians." *The Journal of Politics*. https://doi.org/10.1086/736358.
Barari, Soubhik, Christopher Lucas, and Kevin Munger. 2025. "Political Deepfakes Are as Credible as Other Fake Media and (Sometimes) Real Media." *The Journal of Politics* 87 (2): 510–26.
Blair, Graeme, Jasper Cooper, Alexander Coppock, Macartan Humphreys, and Luke Sonnet. 2025. "Estimatr: Fast Estimators for Design-Based Inference." R package version 1.0.4. https://declaredesign.org/r/estimatr/.
Broockman, David E., and Joshua L. Kalla. 2025. "Selective Exposure and Echo Chambers in Partisan Television Consumption: Evidence from Linked Viewership, Administrative, and Survey Data." *American Journal of Political Science* 69 (3): 847–65.
Culpepper, Pepper D., Jae-Hee Jung, and Taeku Lee. 2024. "Banklash: How Media Coverage of Bank Scandals Moves Mass Preferences on Financial Regulation." *American Journal of Political Science* 68 (2): 427–44.
De Kadt, Daniel. 2025. "Replication of "Instrumentally Inclusive: The Political Psychology of Homonationalism" (Turnbull-Dugarte and López Ortega 2024)." *American Political Science Review*. https://doi.org/10.1017/S0003055425101330.
Franco, Annie, Neil Malhotra, Gabor Simonovits, and L. J. Zigerell. 2017. "Developing Standards for Post-hoc Weighting in Population-Based Survey Experiments." *Journal of Experimental Political Science* 4 (2): 161–72.
Gelman, Andrew, and Hal Stern. 2006. "The Difference between "Significant" and "Not Significant" Is Not Itself Statistically Significant." *The American Statistician* 60 (4): 328–31.
Grose, Christian R., Neil Malhotra, and Robert Parks Van Houweling. 2015. "Explaining Explanations: How Legislators Explain Their Policy Positions and How Citizens React." *American Journal of Political Science* 59 (3): 724–43.
Hainmueller, Jens, and Daniel J. Hopkins. 2015. "The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants." *American Journal of Political Science* 59 (3): 529–48.
Long, Jacob A. 2024. "jtools: Analysis and Presentation of Social Scientific Data." *Journal of Open Source Software* 9(101): 6610.
López Ortega, Alberto, and Stuart J. Turnbull-Dugarte. 2025. "Selectively (Il)Liberal: Theory & Evidence on Nativist Disidentification." *Political Science Research and Methods*. https://doi.org/10.1017/psrm.2025.10066.
Mainz, Sophie. 2026. "Inclusion to Exclude: How Femonationalism Impacts Policy Preferences." *Political Science Research and Methods*. https://doi.org/10.1017/psrm.2025.10068.
Marsh, Wayde Z. C., and Jordan Carr Peterson. 2026. "Public Perceptions of Corporate Position-Taking on Abortion and Transgender Rights." *The Journal of Politics* 88 (1): 401–6.
Matanock, Aila M., and Miguel García-Sánchez. 2018. "Does Counterinsurgent Success Match Social Support? Evidence from a Survey Experiment in Colombia." *The Journal of Politics* 80 (3): 800–14.
McDermott, Grant. 2022. "Ritest: Randomisation Inference Testing." R package version 0.1.0. https://github.com/grantmcdermott/ritest.
Mikkelborg, Anna Caroline. 2025. "White Democrats' Growing Support for Black Politicians in the Era of the "Great Awokening."" *American Political Science Review* 119 (4): 1902–20.
Miratrix, Luke W., Jasjeet S. Sekhon, Alexander G. Theodoridis, and Luis F. Campos. 2018. "Worth Weighting? How to Think about and Use Weights in Survey Experiments." *Political Analysis* 26 (3): 275–91.
Pasek, Josh. 2018. "Anesrake: ANES Raking Implementation." R package version 0.80. https://CRAN.R-project.org/package=anesrake.
Raffaelli, Francesco. 2025. "Homonationalist Rhetoric and Cultural Attitudes." Paper presented at European Political Science Association, Madrid, Spain.
Rainey, Carlisle. 2014. "Arguing for a Negligible Effect." *American Journal of Political Science* 58 (4): 1083–91.
Schaeffer, Merlin, Krzysztof Krakowski, and Asmus Leth Olsen. 2024. "Correcting Misperceptions about Ethno-Racial Discrimination: The Limits of Evidence-Based Awareness Raising to Promote Support for Equal-Treatment Policies." *American Journal of Political Science* 70 (1): 54–75.
Turnbull-Dugarte, Stuart J., and Alberto López Ortega. 2024. "*Instrumentally* Inclusive: The Political Psychology of Homonationalism." *American Political Science Review* 118 (3): 1360–78.
Turnbull-Dugarte, Stuart J., and Alberto López Ortega. 2026. "Replication Data for: *Still* Instrumentally Inclusive." Harvard Dataverse. Dataset. https://doi.org/10.7910/DVN/EWMIKZ.
Wittenberg, Chloe, Matthew A. Baum, Adam J. Berinsky, Justin De Benedictis-Kessner, and Teppei Yamamoto. 2023. "Media Measurement Matters: Estimating the Persuasive Effects of Partisan Media with Survey and Behavioral Data." *The Journal of Politics* 85 (4): 1275–90.

https://doi.org/10.1017/S0003055426101828 Published online by Cambridge University Press