Replication of 'Instrumentally Inclusive: The Political Psychology of Homonationalism' (Turnbull-Dugarte and López Ortega, 2024)*

Daniel de Kadt†

Abstract

Turnbull-Dugarte and López Ortega (2024) argue that increasing exposure to sexually conservative ethnic out-groups causes instrumental support for LGBT+ rights among those pre-disposed to disfavor the ethnic out-group. The paper presents results from two related experiments, one conducted in the UK and a follow up in Spain, where respondents were randomly assigned to read vignettes about anti-LGBT+ protests, and the identity of the protesters was varied. I outline a series of concerns with the paper, primarily related to *ad hoc* empirical choices. The authors use weights (of undisclosed origin) that follow a peculiar bimodal distribution in the second study but use no weights in the first, and inconsistently use robust standard errors throughout. These choices create a pattern of statistically significant results consistent with their theory, a pattern that disappears when either choice is varied. Additional analyses show that, rather than supporting their theory, the second experiment contradicts it.

*I am very grateful to Sarah Brierley, J. Andrew Harris, Wim Louw, Arthur Spirling, and Anton Strezhnev for thoughtful conversations and guidance. I also thank two anonymous reviewers at the *APSR* for their helpful comments. Research documentation and data that support the findings of this study are openly available at the *APSR* Dataverse: https://doi.org/10.7910/DVN/WSZH4H.

†Assistant Professor, Department of Methodology, London School of Economics. Email: d.n.de-kadtlse.ac.uk.

Introduction

Turnbull-Dugarte and López Ortega (2024) consider whether the increasing acceptance of homosexuality in Western countries may be partly attributable to increasing exposure to sexually conservative ethnic out-groups. The authors' theory suggests that such exposure may drive an instrumental increase in LGBT+ tolerance and inclusion by those who are pre-disposed to disfavor the ethnic out-group, termed "homonationalism." The paper presents results from two related experiments, one conducted in the UK ("study 1" in the paper) and one in Spain ("study 2" in the paper). In these two experiments respondents were randomly assigned to read vignettes about protests against LGBT+ education in schools.

In study 1, the control vignette made mention of protesters with conventional white-British names, while the treatment vignette mentioned protesters with identifiably Muslim names and a photograph of protesters in identifiably Muslim dress. In study 2 the treatment vignette likewise featured Muslim names, Muslim organizations, and a photograph of people in identifiably Muslim dress. Post-treatment, the authors measure support for LGBT+ education in schools as their primary dependent variable.

In both the UK and Spain Turnbull-Dugarte and López Ortega (2024) report that being primed with the ethnic out-group (Muslims in both cases) leads to an increase in support for LGBT+ inclusion in schools. Critically, this effect is generally stronger (and in study 1 only present) among those with pre-existing negative attitudes towards immigrants. The authors argue that this heterogeneity in treatment effects is evidence of instrumentalism. Consistent with their theory, individuals who are pre-disposed to disfavor the out-group are more likely to support LGBT+ inclusion in schools when they see that support as instrumental opposition to the disfavored ethnic out-group.

This note outlines a series of concerns with the published version of the paper. The primary concerns relate to seemingly idiosyncratic and *ad hoc* choices made by the researchers in analyzing their data. First, in study 2 the authors elect to use weights (of undisclosed origin) that follow a peculiar bimodal distribution in their regression analyses. Roughly two-thirds of respondents receiving a weight of less than 0.01, and one-third receiving the same weight of approximately 3. No weights are used in study 1. Second, the authors selectively use heteroskedasticity-robust standard errors in both studies. Together, these choices drive the pattern of statistically significant results reported in study 2. When re-analysed, either with no weights, robust standard errors, or both, the results from study 2 do not appear to corroborate the authors' theoretical predictions, or the results found in study 1.

In fact, the results from these re-analyses directly cut against the authors' theoretical prediction that those who are predisposed to be anti-immigrant should be more strongly affected by the treatment. Principally, the assigned weight appears to predict treatment effect heterogeneity, irrespective of the respondents' pre-treatment immigration sentiment. Among those assigned low weights there is neither a first-order effect of treatment nor the hypothesized

interaction effect, despite the fact that many of these respondents report high anti-immigration sentiment. Likewise, among the 372 respondents assigned a high weight in study 2 there is always a first-order effect and interaction effect, even among those with low anti-immigration sentiment. These findings, and other analyses presented throughout, directly contradict the authors' theoretical predictions.

I further document a series of issues related to inconsistencies between the reporting in the text of the paper and the presentation of results, and multiple misleading features of many of the visualizations in the published paper. In sum, Turnbull-Dugarte and López Ortega (2024) make multiple seemingly arbitrary choices that produce a misleading pattern of results, both statistically and visually. Probing these choices causes the results to evaporate, and even directly contradicts the authors' theoretical predictions.

Analytical Inconsistencies

The experimental designs used in study 1 and study 2 are very similar, though the survey platforms used are different, as are the samples. In study 1, the authors use data from a "representative" (p. 1366) sample of 1151 respondents from the UK (details on the origin of the sample, e.g. vendor, population targets, and so forth are not provided in the paper or published supplementary materials). In study 2, the authors use a "crowd-sourced [...] sample" (p. 1370) of 1216 Spanish respondents through the vendor Prolific. Throughout the paper the authors vary their analytical approach to these two studies in *ad hoc* fashion: first, through weighting, and second, in terms of standard error estimation.

The published results of study 2 are based on the inclusion of survey weights in the regression analyses. This is inconsistent with the analysis of study 1, and this inconsistency is not explained or justified in the paper beyond the indication that the data were from a crowd-sourced sample and not a representative sample. Of course, even though the study 1 data are ostensibly a representative sample, there is no reason why one could not use weights in that case too.

The weights themselves are ostensibly post-stratification survey weights designed to "approximate population parameters based on gender, age, education, and geographical region" (p. 1370), yet they follow a highly unusual bimodal distribution, as shown in 1. Roughly two-thirds (63%) of the respondents receive very low weights close to zero (under 0.01), and roughly one third (31%) receive a weight of 3.00009 (rounded to the fifth decimal). Only 6% of observations receive weights greater than 0.01 and less than 3. No information about the process or parameters that generated these weights is provided in the paper, supplementary material, or replication material, and the weights were seemingly not provided by the vendor Prolific (this vendor does not ordinarily provide weights).

When analyzing survey data, weighting choices should depend on the inferential target (estimand) of interest. In survey experiments, weights may be appropriate when it is important to target the population average treatment effect

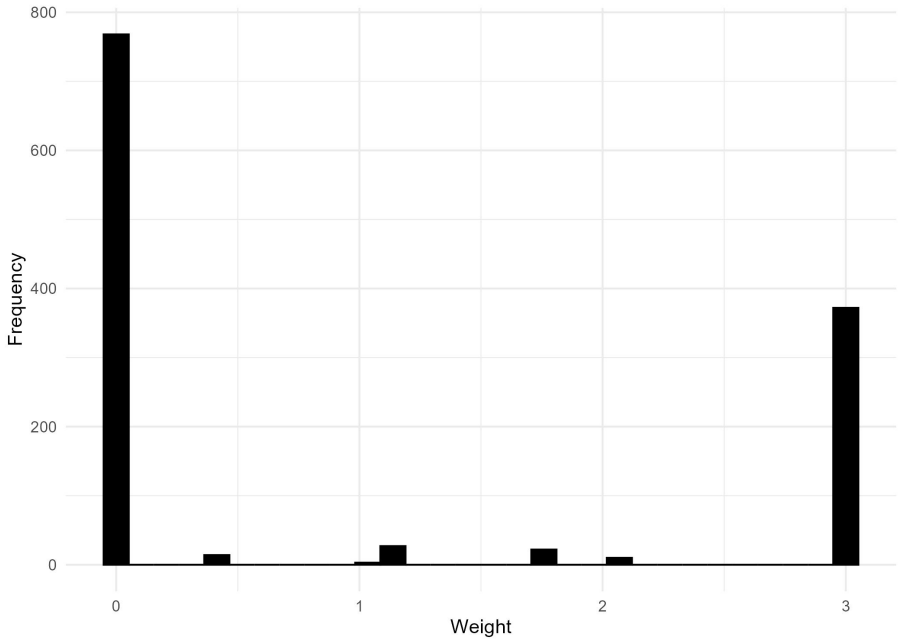

Figure 1: Distribution of Weights in Study 2 (Spain)

Note: This figure shows the distribution of the provided weight variable, for all respondents in the Spain sample.

(PATE) rather than a sample average treatment effect (SATE). Whether these two estimands are very different is impossible to know *a priori*, yet Mullinix et al. (2015) find that, in the case of multiple survey experiments conducted with both unweighted convenience (MTurk) and weighted representative samples (TESS), estimates of the SATE (MTurk) and PATE (TESS) tend to track quite closely. Likewise, using fixed high quality online samples from YouGov, Miratrix et al. (2018) find that weighted and unweighted analyses also tend to be similar.

Miratrix et al. (2018) note that "it is important to compare the PATE and SATE estimates," and warn that when these estimates meaningfully diverge this is "a flag that weight misspecification could be a real concern" (p. 289). The risk of misspecification of the weights is particularly important because if the goal is to estimate the PATE, then the weights themselves should plausibly recover population-level features. Given the absence of any information about the weights and their highly unusual distribution of the study 2 weights, this seems unlikely.

The scholarly consensus in political science appears to be that, if using weights, both the weighted and unweighted estimates should be reported and contrasted, and the "construction and application of weights [should be done] in a detailed and transparent manner" (p. 161 Franco et al., 2017). Neither practice is followed in Turnbull-Dugarte and López Ortega (2024).

A second major inconsistency is that the authors elect to use heteroskedasticity-robust standard errors in some analyses but not in others, as summarized in 1. The table also summarizes a generally *ad hoc* approach to the presentation of confidence intervals, to which I return in detail later in the note. Robust standard errors should likely be used across the board in the paper, but are particularly important in the analyses with binary dependent variables (e.g. tables A7 and A9) as these linear probability models necessarily suffer from heteroskedasticity (Mullahy, 1990). Again the authors do not explain their choices, and again they are not consistent with best practices.

Furthermore, classical standard errors for weighted least squares assume that the weights are precision weights. However, when using sampling or survey (e.g post-stratification) weights, the variance estimator should account for the randomness that stems from the sampling process as reflected in the weights (Lumley and Scott, 2017). Generally, the standard errors for weighted least squares with survey weights will increase compared to the standard error with precision weights when the weights themselves have a high variance, as is the case in study 2.

Re-Analysis of Study 2

Turnbull-Dugarte and López Ortega (2024) conducted study 2 "to assess the external validity of the primary findings [...]" and "[...] to expand upon [those] findings by [asking whether] ethnic out-group opposition result in increased national pride in liberal 'Western' values" (p. 1370). Given the analytical inconsistencies outlined above, I conduct a series of analyses that probe the

Table 1: Summary of Variance Estimation Throughout Paper and Supplementary Materials

	Figure/Table	Robust SEs?	Notes
Study 1:			
	Figure 3	No	90% CIs shown
	Figure 4	Yes	Text overlay uses classical SE, 90% CIs shown
	Figure A4	No	90% CIs shown
	Table A7	No	
	Table A8	Yes	
Study 2:			
	Figure 6	No	90% CIs shown
	Figure 7	No	Text overlay uses classical SE, 90% CIs shown
	Figure 8	No	90% CIs shown
	Figure A5	No	90% CIs shown
	Table A9	No	
	Table A10	No	
	Table A11	Yes	

consequences of the authors' *ad hoc* choices. Study 2 does not support the authors' stated conclusions.

For the sake of transparency I first reproduce the original table A9 from the published paper's supplementary materials, which presents the key results for study 2, as table SM1 in this note's Supplementary Materials (SM). This includes four different analyses where the dependent variable is support for LGBT+ education in schools and the treatment is a binary variable for the vignette condition (1 if treatment, 0 if control):

1. Base model: Estimates the effect of the treatment, conditional on immigration sentiment.
2. Interaction: Estimates the effect of the treatment, allowing the effect to vary linearly by immigration sentiment (higher value means more positive).
3. ProImmigration: Estimates the effect of the treatment, only for the subset of individuals whose immigration sentiment is higher than the mean (in study 2 this is 7 or higher).
4. AntiImmigration: Estimates the effect of the treatment, only for the subset of individuals whose immigration sentiment is lower than the mean (in study 2 this is 6 or lower).

I estimate these four tests with various combinations of weighting and standard error choices. 2 provides an overview of how these choices alter the statistical conclusions presented in the paper. The table summarizes the key coefficients of interest in the regression, their corresponding standard errors, and statistical significance at conventional levels (0.1, 0.05, and 0.01). Results are shown for the four key tests in separate columns – the baseline test, the interaction test, the pro-immigration sample, and the anti-immigration sample. Note that for three of the four tests there is only one coefficient of interest (treatment), while for the interaction test there are two coefficients of interest (treatment and the interaction term, shown in that order).

For each test, there are five possible specifications. The first row ("Weighted + Classical SE") is a direct replication of the original analysis in the paper (corresponding to table SM1 in the SM). To estimate robust standard errors I use the `HC3` option in `estimatr::lm_robust()`, and to estimate survey-robust standard errors I use `survey::svyglm()`. Rows two through five show various combinations of weighting and standard error choices, corresponding to tables SM2, SM3, SM4, and SM5 in the SM.

Table 2: Summary of Researcher Choices, Point Estimates, and Statistical Conclusions

Researcher Choices	Base Model	Interaction	Pro-immig.	Anti-immig.
Weighted + Classical SE	0.095***	0.211*** | -0.019**	0.111***	0.103**
	(0.025)	(0.062) | (0.009)	(0.027)	(0.044)
Unweighted + Classical SE	0.026	0.094 | -0.01	0.037	0.034
	(0.023)	(0.065) | (0.009)	(0.024)	(0.043)
Weighted + Robust SE	0.095**	0.211* | -0.019	0.111**	0.103
	(0.042)	(0.112) | (0.016)	(0.051)	(0.069)
Unweighted + Robust SE	0.026	0.094 | -0.01	0.037	0.034
	(0.023)	(0.075) | (0.01)	(0.024)	(0.043)
Weighted + Survey-R SE	0.095**	0.211* | -0.019	0.111**	0.103
	(0.042)	(0.11) | (0.015)	(0.05)	(0.068)

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 2 shows that the results for study 2 are highly sensitive to researcher choices in terms of both magnitude and statistical significance. It is only when weights are used and classical standard errors estimated that a pattern of statistically significant results consistent with the authors' theory emerges.

The use of weights materially changes the magnitude of the point estimates – these attenuate toward zero by around 1/2× to 1/3× when the weights are removed. Likewise, when using more appropriate variance estimators the standard errors increase by around 2× to 3× across-the-board. It is quite striking that the weighted analyses in study 2 produce point estimates so close to the unweighted analyses in study 1. With weights applied to study 2 (but not to

study 1), the interaction estimate is 0.019 in both studies, and the effect in the anti-immigration sub-sample is approximately 0.1 in both studies.

This fragility is not merely a lack of 'robustness' – the results directly contradict the authors' theoretical argument. First, 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. This is not only a question of statistical inference; 2 shows that without weights the magnitude of the interaction term is halved and substantively close to zero. Second, in all alternative specifications there is no statistically significant result for specifically the anti-immigration sample, the very sample for which the authors' theory predicts an effect would most likely emerge. Again, 2 shows that when weights are removed the estimated treatment effect for this sub-sample is one-third the magnitude of the original estimate, and substantively close to zero.

Similar conclusions can be drawn for the ancillary analysis of Western liberal values. In the SM I reproduce the published table A11 (which includes robust standard errors in the published paper) from the published paper's supplementary materials as table SM6. Without weights there is again neither a statistically significant first-order effect of treatment, nor a statistically significant interaction effect. Substantively, both point estimates attenuate toward zero and are of approximately the same magnitude as many of the "placebo" estimates included in the same table. In sum, study 2 does not support the authors' theory.

Heterogeneous Effects by Weights

The sensitivity to the use of weights is likely exacerbated by their unusual bimodal distribution. This has implications for both the differences in the point estimates between the unweighted and weighted analyses (the estimates are prone to change a lot), and for the standard errors of the estimates (the standard errors are prone to increase due to the relatively high variance of the weights). To probe this I segment the data into three bins – one for those with low weights under 0.01, one for those with high weights greater than or equal to 3, and one for those with weights in-between. The exact choice of the middle bin is arbitrary, but given the bimodal distribution of the weights it matters little.

I focus on the anti-immigration and pro-immigration sub-samples in turn. Table 3 shows that even among those with high anti-immigrant sentiment, the treatment effect is only non-zero and statistically significant for those with high weights. The medium bin captures only a handful of observations so the result there can likely be set aside, but the point estimate for those in the low bin – which captures almost three-fifths of the data in the anti-immigration sub-sample – is essentially zero. The exact same pattern emerges, with point estimates that are very similar, in table 4, despite these respondents being those for whom the authors' theory predicts minimal treatment effects.

A similar analysis using all the data from study 2, presented in table SM7 in the SM, reveals that in none of the weight bins is the estimated interaction term

Table 3: Study 2 (Spain): Heterogeneous Effects By Weight Bin for Anti-Immigrant Sub-Sample

	Low Weights	Mid Weights	High Weights
Treatment	-0.018	-0.182	0.132*
	(0.055)	(0.169)	(0.074)
Intercept	0.687***	0.682***	0.463***
	(0.041)	(0.104)	(0.052)
Num.Obs.	294	38	184
R2	0.000	0.034	0.018
Weighted	No	No	No
HC3 Robust SEs	Yes	Yes	Yes
Survey Robust SEs	No	No	No

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 4: Study 2 (Spain): Heterogeneous Effects By Weight Bin for Pro-Immigrant Sub-Sample

	Low Weights	Mid Weights	High Weights
Treatment	0.007	-0.006	0.128**
	(0.027)	(0.077)	(0.055)
Intercept	0.904***	0.950***	0.766***
	(0.019)	(0.051)	(0.044)
Num.Obs.	474	38	188
R2	0.000	0.000	0.029
Weighted	No	No	No
HC3 Robust SEs	Yes	Yes	Yes
Survey Robust SEs	No	No	No

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

statistically significant. Indeed, it is only in the mid and high weight bins that the interaction term is non-zero. In the low-weight bin, the interaction term is essentially zero. Together, these results again provide evidence contrary to the authors' hypothesis that the treatment effect is conditional on immigration attitudes. Any conditionality in study 2 is in fact with respect to the weights, and not immigration attitudes.

Who is Being Up-Weighted?

The authors state in the published paper that they devised the weights to match Spanish population parameters in terms of gender, age, education, and region. Perhaps it is the case that the weights happen to target those with high levels of anti-immigration sentiment, which would explain the heterogeneous effects in a way consistent with the authors' theory. In table 5 I assess which characteristics are being up-weighted by examining the characteristics of a range of covariates across the weight bins.

Essentially, the weight distribution appears to be driven by age. The Prolific sample skews young compared to the Spanish population, and so those older respondents receive high weights while those who are younger (the bulk of the data) are down-weighted. Respondents in the high weight bin are much older, are much more likely to have children, and are far less likely to identify as queer. Gender, foreign born status, and education appear reasonably well balanced across the bins. Importantly, those in the high weight bin are only weakly more likely to be anti-immigration than those in the low weight bin, further evidence that the dependence of the treatment on weights has little to do with anti-immigration sentiment. I present an in-depth analysis of treatment effects by age group in both the UK and Spain study in the SM, in tables SM8 and SM9, and Figures SM1, SM2, and SM3.

Table 5: Study 2 (Spain): Covariates by Weight Bin

Variable	Low Weight Bin	High Weight Bin	Medium Weight Bin
age	26.17	39.64	31.11
gender	0.48	0.51	0.54
edu	3.27	3.45	3.57
child	0.07	0.37	0.2
foreignborn	0.21	0.25	0.21
queer	0.3	0.16	0.3
imm_1	6.9	6.16	6.36

Visual and Reporting Inconsistencies

As outlined above, Turnbull-Dugarte and López Ortega (2024) propose heterogeneous treatment effects that increase with pre-treatment anti-immigration attitudes as the dispositive evidence of their theory. In both study 1 and study 2 this heterogeneity is presented through two primary visualization techniques, while regression tables are relegated to the supplementary material. The first set of published visualizations, figures 3 (UK) and 6 (Spain) respectively, are interaction plots that purport to show how the treatment effect varies across the span of immigration sentiment. The second set, figures 4 (UK) and 7 (Spain) respectively, are dot-plots showing the means and confidence intervals for treatment and control across two subgroups, those who are pro-immigration (defined as higher than the mean), and those who are anti-immigration (lower than the mean). Both sets of visualizations are misleading.

Figures 3 and 6: The Interaction Plots

In the body of the paper the authors state that they "estimate the CATE via [...] a linear estimation of the conditionality of moderator values" using the following specification (equation 1, p. 1367):

$$Y_i = \alpha + \delta_1 treat + \beta_1 imm_1 + \beta_2 treat * imm_1 + \epsilon_i$$

Given the experimental design, the variable `treat` represents the vignette treatment and `imm_1`, an 11-point scale variable, the moderator. Note that there is likely a typographical error in the equation in the published paper (which I have corrected here), as one would typically model an interaction as a new parameter, rather than the product of both the parameters and the variables. However, it is clear in the text and the formalization that a linear regression with an interaction term underpins the analysis.

Figures 3 and 6 in the published paper purport to summarize the results of this analysis, and the authors direct readers to supplementary material tables A7 and A9 respectively for "full regression output" in the figure notes. Each figure includes two different panels which convey roughly the same information. The top panel shows how the predicted value of Y changes as a function of immigration attitudes, for each level of the treatment variable – there are thus two curves, one in dashed red (for the treated group) and one in solid blue (for the control group). The reader is asked to interpret the gap between the curves as the conditional average treatment effect for a given level of `imm_1`. The bottom panel of the figures presents that difference as a point for each discrete level of the moderating variable, with confidence intervals around each point.

Neither figure is based on the stated empirical specification or the regression tables to which readers are directed. Inspecting the replication code, the figures are instead generated with the `jtools::interact_plot()` function in `R` on the basis of an underlying logistic regression, not a linear regression. This discrepancy between the stated regression specification and the actual implementation is never noted in the paper or in the figures. The published tables A7 and

A9 to which readers are directed present the results of linear regressions, not the logistic regressions that underpin the figures. While logistic regression is a reasonable (and perhaps even commendable) choice for interacted specifications with binary dependent variables, the discrepancy between the text, the figures, and the supplementary tables remains.

The visible non-linearity in both panels in the published figures 3 and 6 is thus a function of the use of logistic regression. Heterogeneous treatment effects are not separately estimated with a fully-saturated regression including first-order and interaction terms for each level of the moderator. Only four parameters are estimated: the intercept (where the treatment and immigration are both 0), the treatment effect (the mean shift in Y for a change in treatment status), the first order coefficient for immigration (how Y shifts for a 1 unit change in immigration sentiment), and the interaction term (any additional shift in Y for a 1 unit change in immigration sentiment for those who are treated only). Because these analytical choices are not explained in the paper, and because the authors do not indicate that their specification is a logistic regression, the visible nonlinearities in the published Figures 3 and 6 are misleading – readers may erroneously believe that the treatment effect is estimated for each level of the moderator, or that there is some more complex heterogeneity across the span of `imm_1`.

Somewhat to this point, the authors report the results of a "linearity test" styled after Hainmueller et al. (2019) in their supplementary materials (the published figures A6 and A7). Unfortunately these tests are implemented in a confusing fashion using the `jtools::interact_plot()` function. The authors appear to mistakenly switch the predictor value (which should be `treat`) for the moderator value (which should be `imm_1`) – the tests (and the published figures A6 and A7) thus reveal a roughly linear relationship between immigration sentiment and the outcome variable, but do not directly test for linearity in the effect of treatment over the span of the moderator. Neither the purpose of this test as specified, nor the result of this implementation, are explained in the paper or the supplementary material. Fortunately, more appropriate tests of linearity do suggest that linearity is not unreasonable in this case (not reported in this note).

Finally, it is worth contemplating the presentation of uncertainty in the published figures 3 and 6. First, the authors do not include confidence bands for the top panel of the figure – the reason for this omission is unclear. However, as noted earlier, in the lower panel the authors present only 90% confidence intervals based on classical standard errors. This detail is not mentioned anywhere in the paper or the figures, and was only discernible by examining the hard-coded critical values used to generate the confidence intervals in the replication materials. While there is certainly no hard and fast rule about which confidence intervals one should present, and any particular level of α is ultimately an arbitrary choice, 95% confidence intervals are expected and assumed as default and to present 90% intervals with no indication of that fact is misleading.

I reproduce the published figures 3 and 6 below, with two corrections, as figures 2 and 3 respectively. First, I correct the code to use the specification

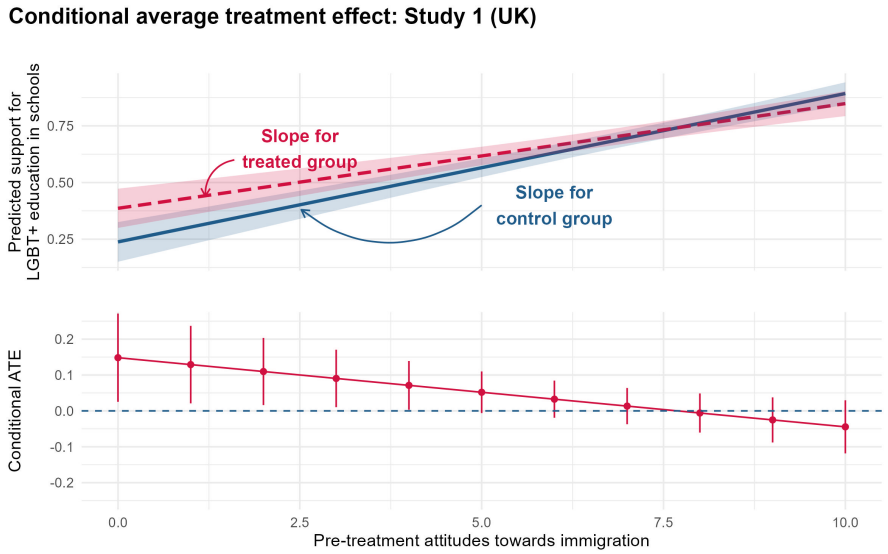

Figure 2: Study 1 (UK): Corrected Reproduction of Figure 3

Note: This figure shows results from an interaction analysis for the UK sample. The top panel shows the predicted support for dichotomized LGBT+ education in schools for the control group (blue) and the treated group (red), across the span of the moderator, pre-treatment attitudes toward immigration (higher values imply higher support for immigration). The bottom panel shows the conditional average treatment effect (CATE) for each level of the moderator. All results come from a linear regression of the outcome on a treatment indicator, the moderator, and the interaction of these two variables. All confidence intervals are 95% confidence intervals.

described in the paper – linear regression estimated with ordinary (weighted) least squares. To reiterate, there is nothing necessarily wrong with logistic regression, but it is not the analysis proposed by the authors in the paper. Second, I correct the confidence intervals to be 95% confidence intervals based on robust standard errors, and include confidence bands based on robust standard errors in the upper panel. I do not reproduce the published Figure 8, but the same issues apply. Additionally in the SM I reproduce the published Figure 6 for study 2 as figure SM6, which includes the above corrections but removes the weights.

Figures 4 and 7: The Dot-Plots

The second set of visualisations, figures 4 and 7 in the published paper, are jittered (noise-applied) dot-plots with group means and confidence intervals, generated with modified code based on `jtools::effect_plot()` in `R`. These figures appear to be styled after Coppock (2021), who advocates for such vi-

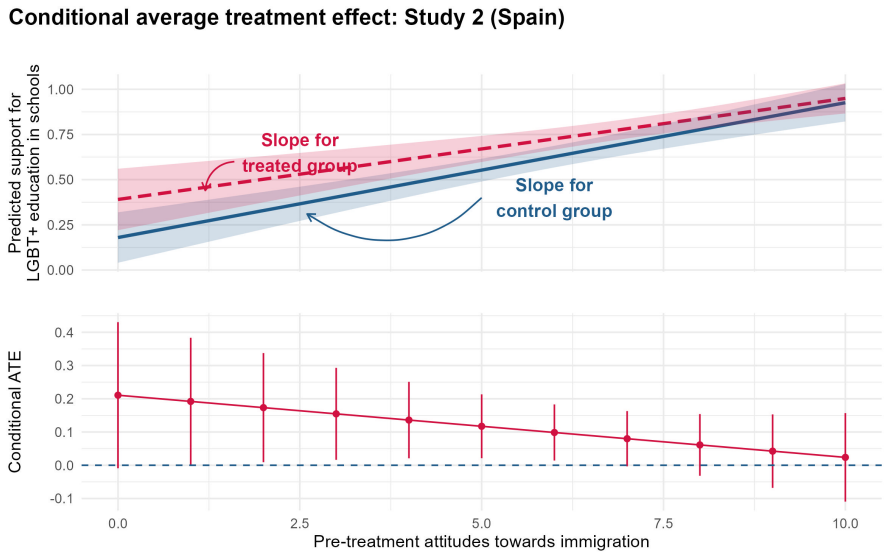

Figure 3: Study 2 (Spain): Corrected Reproduction of Figure 6 (Weights Used)

Note: This figure shows results from an interaction analysis for the Spain sample. The top panel shows the predicted support for dichotomized LGBT+ education in schools for the control group (blue) and the treated group (red), across the span of the moderator, pre-treatment attitudes toward immigration (higher values imply higher support for immigration). The bottom panel shows the conditional average treatment effect (CATE) for each level of the moderator. All results come from a weighted linear regression of the outcome on a treatment indicator, the moderator, and the interaction of these two variables, using the provided weights. All confidence intervals are 95% confidence intervals.

sualizations because they clearly communicate not only the treatment effects of interest, but also the underlying research design that motivates the statistical analysis, the data used to estimate the treatment effect, and statistical uncertainty in any estimates. These figures are again misleading.

The points in the published figures 4 and 7 are not the underlying data on which the analyses are based. Each data point's value on the y-axis is not that data point's value on the binary outcome variable Y or the continuous variable on which the dichotomization was based. Instead they are the fitted value $\hat{Y}_i$ of the binary outcome for each observation, given the results from the underlying regression of the outcome on the treatment indicator. For each of the two subsamples in each experiment (pro-immigration and anti-immigration) there are, of course, only exactly two unique fitted values: one for the treated units and one for the control units. As such, almost all of the visualized variation in the plots comes exclusively from the jittering of these four unique values in 'R'. This is misleading and is not explained in the paper, either in the text or the figures.

Further, the published versions of the plots once again present 90% confidence intervals, a fact documented in neither the figure caption or the paper. Finally, as noted earlier, while in the published figure 4 the confidence bands are based on robust standard errors, the superimposed text-based representation of significance is not. In Figure 7, neither the confidence bands nor the text representation is based on robust standard errors.

I reproduce the published figures 4 and 7 as figures 4 and 5 respectively, with a number of corrections. The figures below show the actual (jittered) value of Y for each observation, and the confidence intervals are the 95% confidence intervals based on robust standard errors. The text representations of point estimates and statistical significance are also updated to reflect estimation with robust standard errors. In study 1 the corrected figures do not much alter the substantive conclusions in the published paper, though the confidence intervals widen and the p-values do increase across the board. In study 2 the changes are more notable.

The updated figures reveal a broader point of concern, most acutely in the Spain sample but also in the UK sample: there are far fewer respondents who take a zero value on the dependent variable in either the treatment or control conditions. In study 1, across both conditions the proportion of the pro-immigrant sub-sample that takes a zero value on the dependent variable is 0.2. In study 2 that proportion is just 0.11.

This raises concerns about potential ceiling effects that might produce artificial heterogeneous effects. If those who are pro-immigration are mostly already pro-LGBT+ education in schools, then there is not much scope for a positive treatment effect – there are few people who can be moved toward favouring LGBT+ education. By contrast, if those who are anti-immigration are more split on LGBT+ education, then there is unsurprisingly more scope for a treatment effect to emerge. This may generate heterogeneity that is observationally equivalent with the authors' theory, when really it is driven by a ceiling effect. Indeed, this provides a useful example of the value of the approach to visualiz-

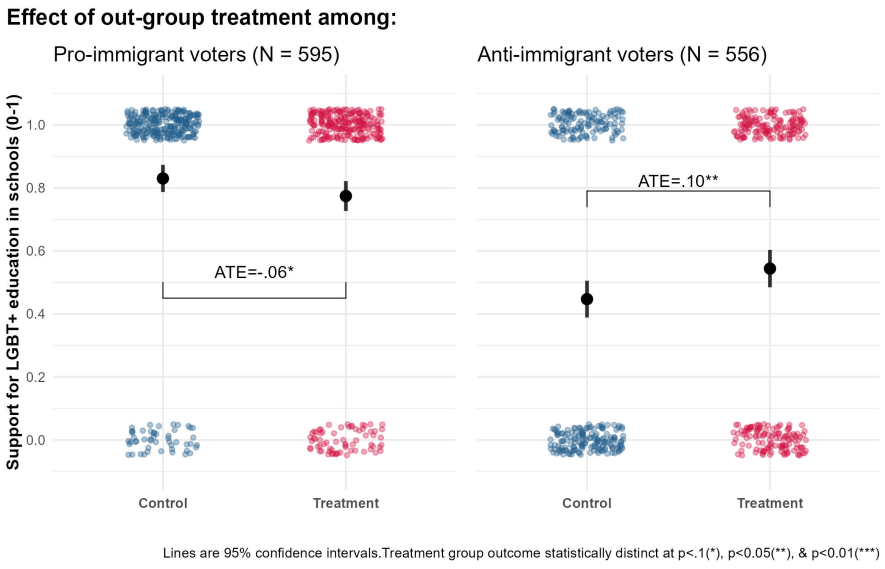

Figure 4: Study 1 (UK): Corrected Reproduction of Figure 4

Note: Each dot represents the value of the dependent variable for each individual respondent in the UK sample, jittered on both axes for visibility, red for treated and blue for control. Black point estimates and overlayed text come from a linear regression of dichotomized support for LGBT+ education in schools on the treatment indicator. Lines represent 95% confidence intervals. The left panel includes only those respondents who scored higher than the mean on immigration sentiment, and the left those who scored lower than the mean.

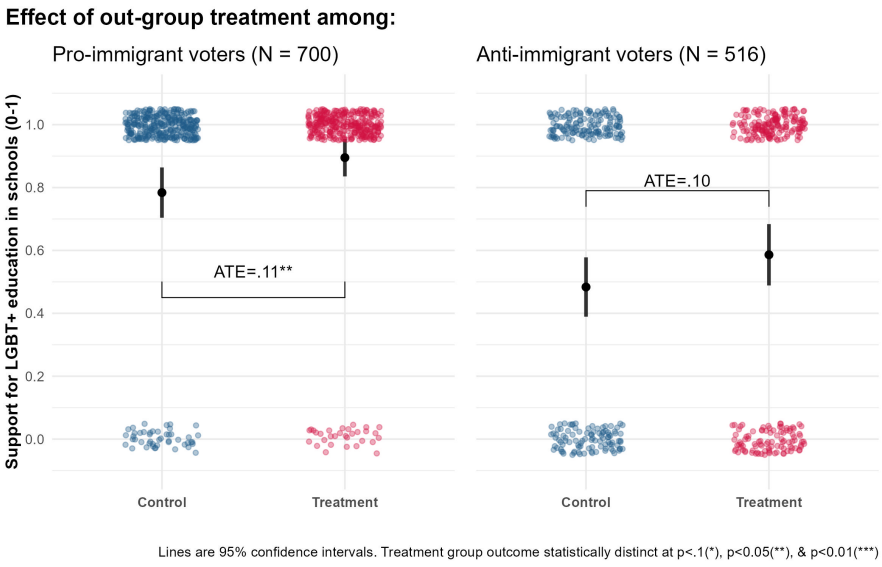

Figure 5: Study 2 (Spain): Corrected Reproduction of Figure 7 (Weights Used)

Note: Each dot represents the value of the dependent variable for each individual respondent in the Spain sample, jittered on both axes for visibility, red for treated and blue for control. Black point estimates and overlayed text come from a linear regression of dichotomized support for LGBT+ education in schools on the treatment indicator, using the provided weights. Lines represent 95% confidence intervals. The left panel includes only those respondents who scored higher than the mean on immigration sentiment, and the left those who scored lower than the mean.

ing treatment effects proposed by Coppock (2021), but this value is of course conditional on the visualizations being appropriately implemented.

Conclusion

Turnbull-Dugarte and López Ortega (2024) argue that increasing support for LGBT+ rights in Western countries is at least in part driven by instrumental "homonationalism." Increasing exposure to sexually conservative ethnic out-groups may drive an instrumental increase in LGBT+ tolerance and inclusion among those who are pre-disposed to disfavor the ethnic out-group. To test this argument the authors conduct two similar online survey experiments, one in the UK (study 1) and a follow up in Spain (study 2). Respondents are presented with a vignette about protests against LGBT+ education, and the treatment condition labels the protesters as identifiably Muslim.

In both studies Turnbull-Dugarte and López Ortega (2024) report that the treatment increases support for LGBT+ inclusive education in schools, and that the effect increases as a function of pre-existing anti-immigrant sentiment. In this note I have highlighted a number of issues with the paper, in turn undermining the support for the authors' theory. Study 2, conducted in Spain, is particularly problematic: the pattern of statistically significant results presented in the paper is driven by idiosyncratic and *ad hoc* choices made by the authors with respect to weights and standard error estimation. Because neither study was pre-registered it is hard to make sense of the various choices made.

While I have not engaged the results of study 1, the visualizations used to present results from that study, as well as study 2, are misleading and do not accurately represent the underlying data or the statistical analyses. For the published figures 3 and 6 this includes a misrepresentation of the underlying statistical models being visualized. For the published figures 4 and 7 this includes the misleading use of jittered predicted values from a regression. Additionally, the figures report only 90% confidence intervals, but do not mention this fact.

There are other minor points of inconsistency throughout the paper. In the UK experiment (published tables A7/A8) the "base model" does not include any covariates, while in the Spain experiment (published tables A9/A10) the base model includes a control for pre-treatment attitudes towards immigration. Likewise, the dichotomizing of the dependent variables is inconsistent, dichotomized at the mean on an 11 point scale for the main dependent variable and at 5 on an 11 point scale for the ancillary dependent variable.

Turnbull-Dugarte and López Ortega (2024) present a theory which they test with two similar survey experiments. For one of those experiments – study 2 – the results are sensitive to seemingly *ad hoc* choices made about weights and standard errors. Further analyses of that experiment reveals evidence that contradicts the authors' theoretical predictions. While the results of study 1 do not appear to be sensitive to the same issues, the presentation of results from both studies is analytically inconsistent and systematically misleading.

There are two potentially general lessons here for applied researchers about

the presentation of results. First, while there has been a move toward focusing on the visual presentation of results, and in particular the use of visualizations to communicate treatment effects, we should be cautious. Regression tables may be less visually appealing but they are also less prone to mislead, and easier for readers to interrogate. Had the published results been presented as tables, instead of visualizations, it is quite possible that the issues I have highlighted would have been more readily apparent without the need to dive into the replication code. Second, when creating visualizations, it is imperative that researchers be transparent and clear about what is being visualized (whether data or model), and provide clear text-based notes that accurately explain for the reader what exactly they are being shown.

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/WSZH4H.

ACKNOWLEDGMENTS

I am very grateful to Sarah Brierley, J. Andrew Harris, Wim Louw, Arthur Spirling, and Anton Strezhnev for thoughtful conversations and guidance. I also thank two anonymous reviewers at the *APSR* for their helpful comments.

CONFLICT OF INTEREST

The author declares no ethical issues or conflicts of interest in this research.

ETHICAL STANDARDS

The author affirms this research did not involve human participants.

References

Coppock, Alexander. 2021. "Visualize as You Randomize". In *Advances in Experimental Political Science*, eds. Jamie N Druckman and Donald P Green, Chapter 17, 320–335. Cambridge: Cambridge University Press.

Franco, Annie, Neil Malhotra, Gabor Simonovits, and LJ Zigerell. 2017. "Developing Standards for Post-Hoc Weighting in Population-Based Survey Experiments". *Journal of Experimental Political Science* 4 (2): 161–172.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu. 2019. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice". *Political Analysis* 27 (2): 163–192.

Lumley, Thomas and Alastair Scott. 2017. "Fitting Regression Models to Survey Data". *Statistical Science*: 265–278.

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): 275–291.

Mullahy, John. 1990. "Weighted Least Squares Estimation of the Linear Probability Model, Revisited". *Economics Letters* 32 (1): 35–41.

Mullinix, Kevin J, Thomas J Leeper, James N Druckman, and Jeremy Freese. 2015. "The Generalizability of Survey Experiments". *Journal of Experimental Political Science* 2 (2): 109–138.

Turnbull-Dugarte, Stuart J and Alberto López Ortega. 2024. "Instrumentally Inclusive: The Political Psychology of Homonationalism". *American Political Science Review* 118 (3): 1360–1378.