A Comment on 'Instrumentally Inclusive: The Political Psychology of Homonationalism' (Turnbull-Dugarte and López Ortega (2024)).

Abstract

Turnbull-Dugarte and López Ortega (2024) argue that increasing exposure to sexually conservative ethnic out-groups causes an instrumental increase in support for LGBT+ rights by those who are 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. This comment outlines a series of concerns with the paper, related to idiosyncratic and ad hoc data analysis choices. The authors use weights (of undisclosed origin) in the second study but not 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 likely contradicts it.

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. 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 protestors with conventional white-British names, while the treatment vignette mentioned protestors with identifiably Muslim names and a photograph of protestors 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 the authors find 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 (in study 1 only present) among those with pre-existing (pre-treatment) 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 that disfavored ethnic out-group.

This comment outlines a series of concerns with the published version of the paper. Two 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 published results for 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 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 some of the visualisations in the published paper. In sum, the 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 and the properties of the samples are thus quite different too. 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 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 use 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 Figure 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).

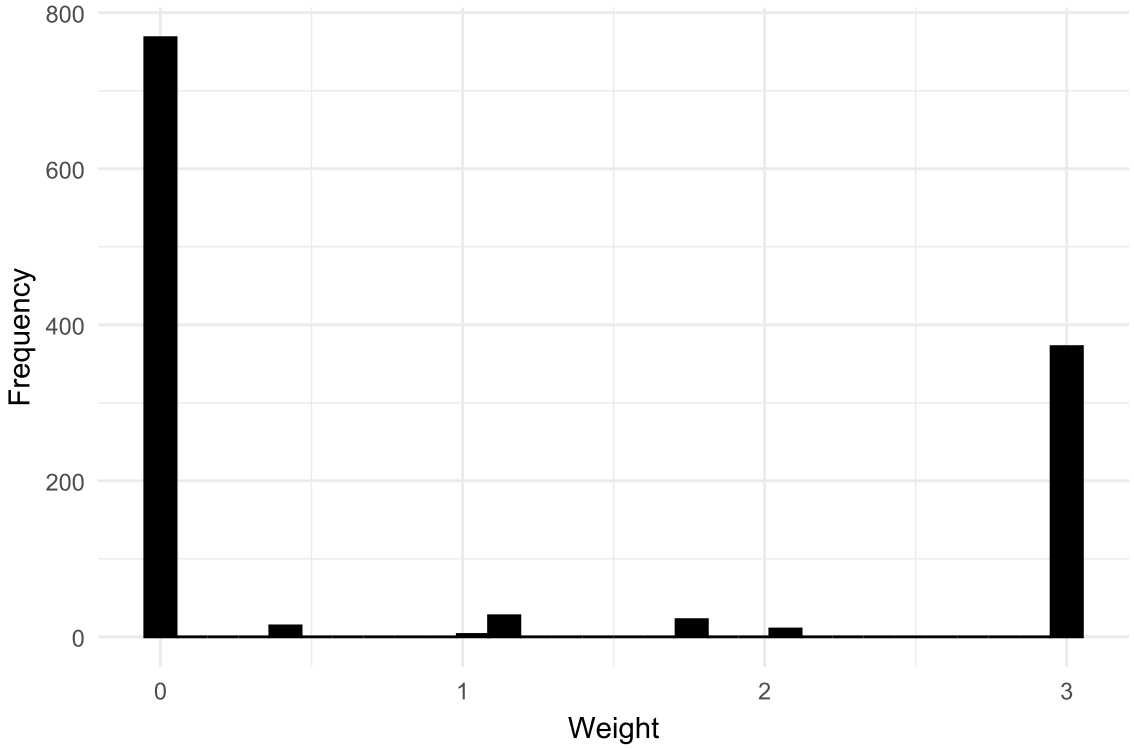

Figure 1: Distribution of Weights in Study 2 (Spain)

A number of papers in political science consider the question of weighting samples for survey experiments Miratrix et al. (2018). Weighting choices depend on the inferential target (estimand) of interest. Weights may be appropriate when it is important to target the population average treatment effect (PATE) rather than a sample average treatment effect (SATE). Whether these two estimands are very different is impossible to know *a priori*. However, 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. Using fixed high quality online samples from YouGov, Miratrix et al. (2018) find that weighted and unweighted analyses also tend to be similar. They note that "it is important to compare the PATE and SATE estimates," and warn that when the estimates meaningfully diverge this is "a flag that weight misspecification could be a real concern" (p. 289).

Importantly, if the goal is to estimate the PATE, then the weights themselves must not be misspecified and should plausibly recover population-level features. Given the highly unusual distribution of the study 2 weights, this seems unlikely. The overriding consensus, then, is that if using weights both the weighted and unweighted estimates should be reported and contrasted, and the "construction and application of weights in a detailed and transparent manner" (Franco et al. 2017, 161). 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 Table 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 comment. 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.

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	

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 for survey weights will increase compared to the standard error for 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 our findings by [asking whether] ethnic outgroup 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 consequences of the authors' *ad hoc* choices with respect to weighting and standard error estimation. I conclude that study 2 does not support these 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 6 in Appendix 2. 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 means 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 means 6 or lower).

I estimate these four core tests with various combinations of weighting and standard error choices. Table 2 provides an overview of how these various 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 – 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, 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 6 in Appendix 2). 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 Table 7, Table 8, Table 9, and Table 10 in Appendix 2.

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)

* p < 0.1, ** p < 0.05, *** p < 0.01. SEs in parentheses.

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 $1/2\times$ to $1/3\times$ when the weights are removed. Likewise, when using more appropriate variance estimators the standard errors increase by around $2\times$ to $3\times$ 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, 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; Table 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, Table 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 Appendix 2 I reproduce Table A11 (which includes robust standard errors in the published paper) from the

published paper's supplementary materials as Table 11. Without weights there is again neither a statistically significant first-order effect of treatment, nor a statistically significant interaction effect. Substantively, both point estimates are also closer to zero and 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. 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 almost exactly the same, in Table 4, despite these respondents being those for whom the authors' theory predicts minimal treatment effects.

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

* p < 0.1, ** p < 0.05, *** p < 0.01

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

* p < 0.1, ** p < 0.05, *** p < 0.01

A similar analysis using all the data from study 2, presented in Table 12 in Appendix 2, reveals that none of the estimated interaction terms in any of the weight bins is 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 respondents that are older 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 Appendix 2.

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	0.20
foreignborn	0.21	0.25	0.21
queer	0.30	0.16	0.30
imm_1	6.90	6.16	6.36

Visual and Reporting Inconsistencies

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. For both study 1 and study 2 this heterogeneity is presented through two primary visualisation techniques in the published paper, while regression tables are presented in the supplementary material. The first set of visualisations, Figures 3 and 6 respectively, are interaction plots that purport to show how the treatment effect varies across the span of immigration sentiment. The second set, Figures 4 and 7 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 visualisations are misleading.

Figures 3 and 6: 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 design, the variable `treat` represents the treatment variable 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 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. The figures are instead generated with the `jtools::interact_plot()` function in `R` on the basis of an underlying regression implemented as `glm(support ~ treat*imm_1, data=[DATA], family="binomial")`. This is not a linear regression, but a logistic regression. This discrepancy between the stated regression specification and the actual implementation is never noted in the paper or in the figures. Tables A7 and A9 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 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 beta 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 non-linearities in Figures 3 and 6 are misleading – they may lead readers to erroneously believe that the treatment effect is estimated for each level of the moderator, or that there is some more complex estimated (linear) 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 the Appendix (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 Figures A6 and A7) 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 either 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 comment).

Finally, it is worth contemplating the presentation of uncertainty in 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 Figures 3 and 6 below, with two corrections, as Figure 2 and Figure 3. First, I correct the code to use the specification described in the paper, that is, linear regression estimated with ordinary least squares. Again, there is nothing necessarily wrong with logistic regression, but it is not the analysis outlined by the authors in the paper. Second, I correct the confidence intervals to be 95% confidence intervals based on robust standard errors, and include the confidence bands based on robust standard errors in the upper panel. I do not reproduce Figure 8, but this figure is also based on logistic regression and the same issues apply. Additionally in Appendix 2 I reproduce the published Figure 6 as Figure 7, which includes the above corrections but removes the weights.

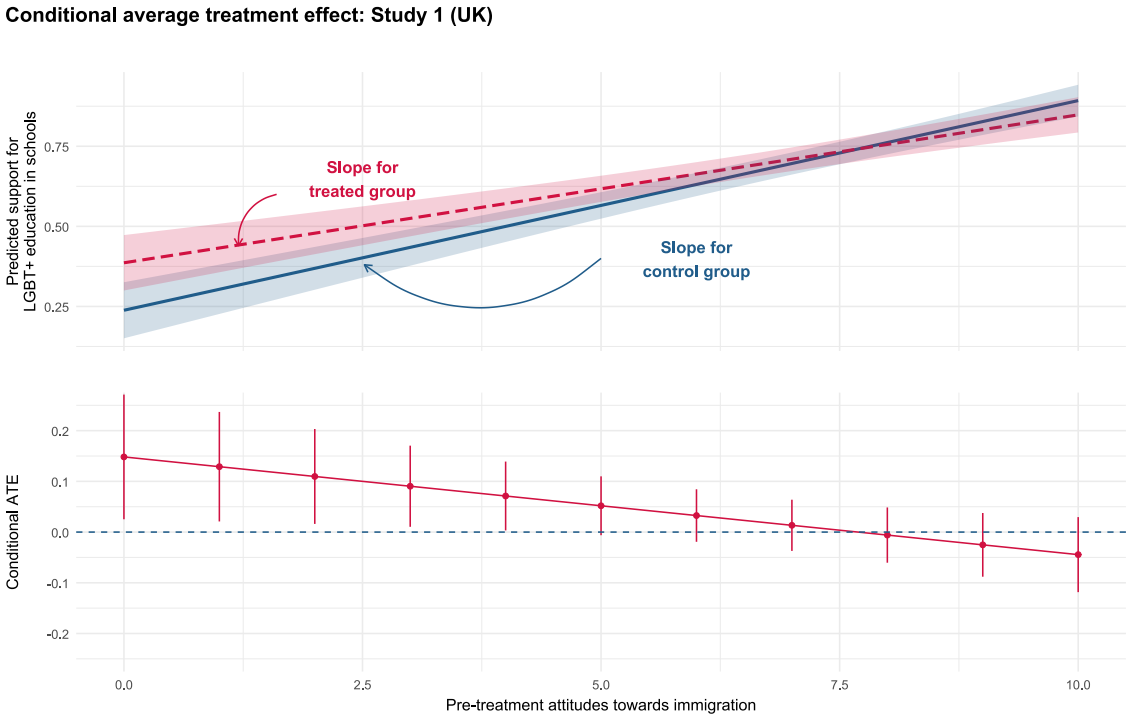

Figure 2: Study 1 (UK): Corrected Reproduction of Figure 3

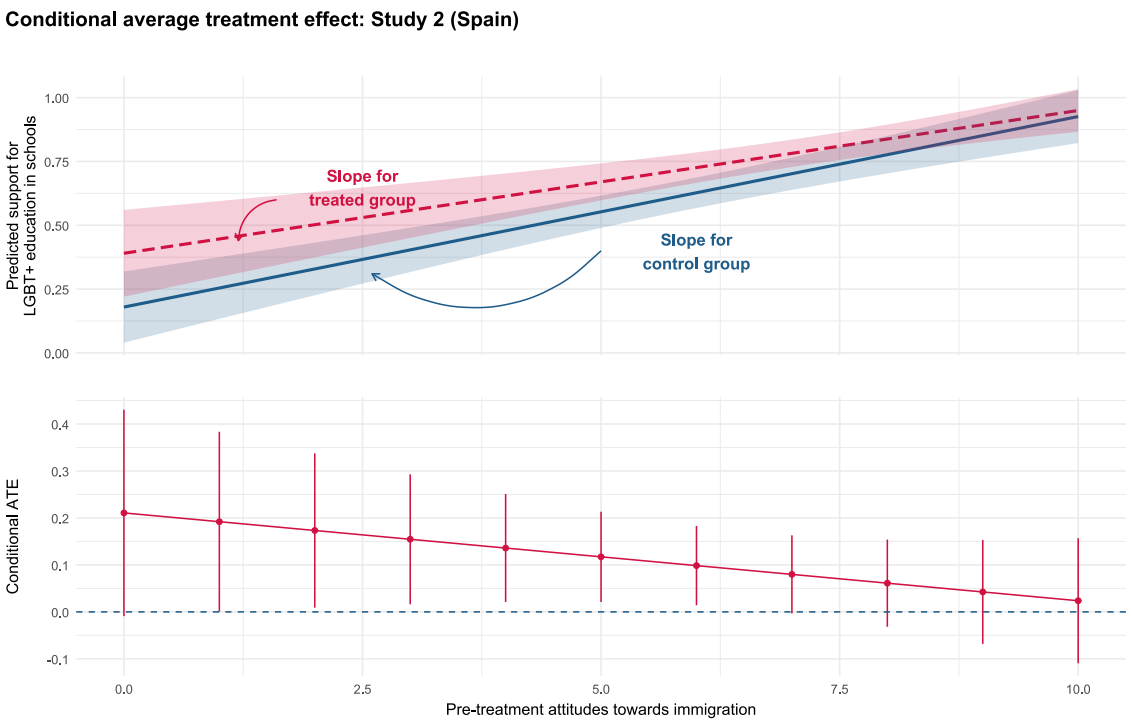

Figure 3: Study 2 (Spain): Corrected Reproduction of Figure 6 (Weights Used)

Figures 4 and 7: The Dot-Plots

The second set of visualisations, Figures 4 and 7 respectively, are jittered (noise-applied) dot-plots with group means and confidence intervals, generated with modified code based on `jtools::effect_plot()` in R. The goal of the plots is to show the means for each condition in each sub-group, the treatment effects for each sub-group comparison, and the underlying data. The figures appear to be styled after Coppock (2021), who advocates for such visualizations because they clearly communicate not only the treatment effects of interest, but also the underlying research design that motivates the statistical analysis, the data that are used to estimate the treatment effect, and statistical uncertainty in any estimates. Unfortunately, these figures are again misleading.

The data points in Figures 4 and 7 do not represent 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 (or "predicted") value $\hat{Y}_i$ of the binary outcome for each observation, given the results from the underlying regression. For each of the two sub-samples 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 visualised 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. Once more, the published versions of the plots only present 90% confidence intervals, but this choice is again not documented in the figure caption or the paper. Finally, as noted before, while in 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 Figures 4 and 7 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 also 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.

In both cases 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. Likewise, 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 suggest 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 Coppock's (2021) approach to visualizing treatment effects, but this value is of course conditional on the visualizations being correctly implemented.

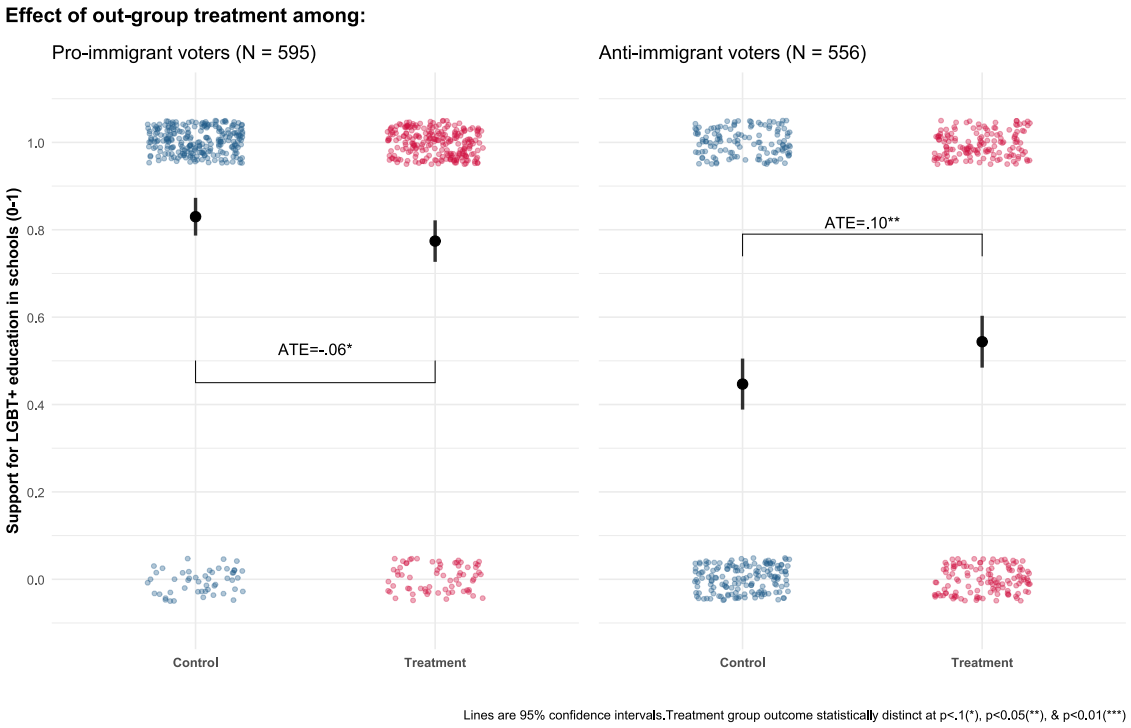

Figure 4: Study 1 (UK): Corrected Reproduction of Figure 4

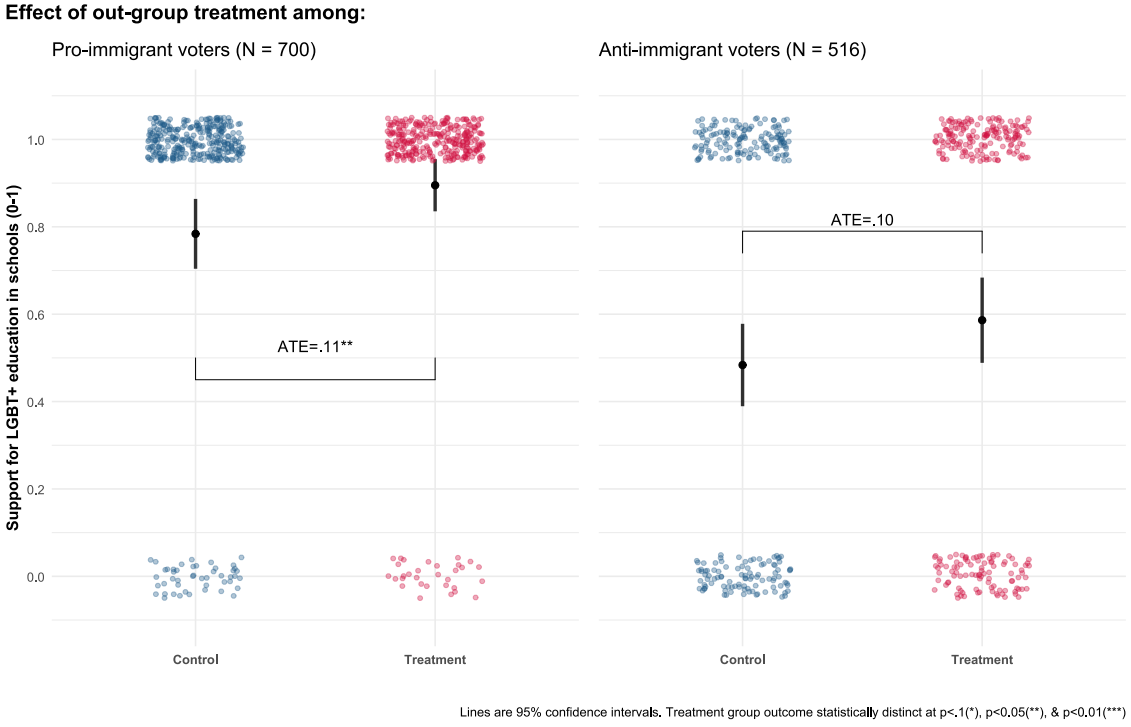

Figure 5: Study 2 (Spain): Corrected Reproduction of Figure 7 (Weights Used)

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). In both studies they report that the treatment increases support for LGBT+ inclusive education in schools, and that the effect increases among those who have higher levels of anti-immigrant sentiment. In this comment I have highlighted a number of issues with the paper, which undermine the support for that theory. Study 2, conducted in Spain, is particularly problematic: the pattern of 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. While I have not engaged the results of study 1, the visualisations 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.

Most pressingly, the results of the Spain experiment rely on the use of post-stratification survey weights in the regression analyses. No such weights are used in the analysis of the UK experiment. This inconsistency is not explained or justified. When the data are not weighted, the results in study 2 are largely null, both statistically and substantively. This sensitivity appears to be driven by the weights' peculiar distribution, with roughly two-thirds of the data receiving weights less than 0.1, and roughly one-third receiving weights of approximately 3. Beyond the choice to use weights, how the weights were created is not sufficiently documented in the paper, supplementary materials, or replication code. Additional analyses suggest that those with high weights appear respond to the treatment, while the bulk of the respondents do not respond in any fashion. Those who have low weights exhibit neither a first-order effect of treatment nor the hypothesized interaction effect, in contrast to the authors' theoretical predictions.

In addition, standard errors throughout the paper are estimated in an *ad hoc* fashion. While all of the analyses should likely use heteroskedasticity-robust standard errors, some do and some do not. This inconsistency is not documented in the paper or the supplementary materials, and has material implications for the statistical significance of the tests in study 2. Because neither study was pre-registered (this is not mentioned in the paper or the supplementary materials) it is very hard to make sense of the choices that seem to be driving the results.

A number of additional issues render the presentation and visualization of results throughout the published paper misleading. For Figures 3 and 6 in the published paper this includes a misrepresentation of the underlying statistical models being visualized. For Figures 4 and 7 in the published paper 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. The specification in Tables A7 and A9 and in Tables A8 and A10 are inconsistent. In the UK experiment (A7/A8) the "base model" does not include any covariates, while in the Spain experiment (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 is potentially a general lesson 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. 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 visualising, 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 they are seeing.

References

Coppock, Alexander. 2021. "Visualize as You Randomize." *Advances in Experimental Political Science*, 320.

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): 161–72.

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): 163–92.

Lumley, Thomas, and Alastair Scott. 2017. "Fitting Regression Models to Survey Data." *Statistical Science*, 265–78.

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.

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): 109–38.

Turnbull-Dugarte, Stuart J, and Alberto López Ortega. 2024. "Instrumentally Inclusive: The Political Psychology of Homonationalism." *American Political Science Review* 118 (3): 1360–78.

Appendix 1: Note on Computational Reproducibility

This appendix reports on computational reproducibility. The paper is almost perfectly computationally reproducible from the processed data provided. That is, the authors provide the code and data required to (almost exactly) reproduce all results in the paper and supplementary materials, and the code runs (largely) without error. However, the data have evidently been pre-processed (e.g. the creation of the weights which were likely not provided by Prolific), and neither the raw data nor this pre-processing code are provided. I note a few minor issues encountered below:

- The `modelsummary::modelsummary` function does not support the use of `robust = TRUE` as an argument for returning robust standard errors. Despite this, the standard errors in the paper are robust where this argument is used (perhaps a prior version supported this). The correct result should be achieved with `vcov = "HC3"` instead.
- The `ritest` package not available on CRAN, and must be downloaded using: `remotes::install_github("grantmcdermott/ritest")`. Ideally this would be noted in the materials.
- The `starbility` package not available on CRAN, and must be downloaded using: `remotes::install_github('https://github.com/AakaashRao/starbility')`. Ideally this would be noted in the materials.
- When using `ggsave()` it would be wise to hard-code the dimensions of the output figure. This function defaults to the current dimensions of the plot window in RStudio (or whatever graphics device is currently active), which will vary by user.
- In the script `study1_summarystats.R` the code erroneously asks for `UKdata_analysis.csv` which is not provided in the replication archive. The correct file is `study1_data.csv` – when correcting this the results are reproduced. There are also some slight inconsistencies in the tables produced by this code and the tables in the supplementary materials (e.g. there are rows returns by the code for Table A.3. that are not in the supplementary materials). There is a similar discrepancy in the code that produces Table A.5. in `study2_summarstats.R`. In this file there are two lines that seem to produce summary statistics tables (one that is weighted, one that is not), but only one is in the supplementary materials (the weighted one). This is not indicated in the table.
- For the power analyses, while I was able to reproduce the results in the supplementary information by running the provided code, I was not able to trace the input values myself. These values are hard-coded in the replication code, but it is unclear where the values come from (they do not appear to come directly from the replication data, based on my cursory explorations).
- It is worth noting that the replication materials do not include any data processing code. Though this is not generally required by journals at present and is very rare in the discipline, it would be particularly useful for understanding some of the issues with regards to the weights in the Spain experiment.

Appendix 2: Regression Results and Additional Analyses

This appendix includes a series of reproductions and analyses of the authors' data, with a primary focus on study 2.

Re-analyses of main specifications

In this subsection I report the primary re-analyses of study 2, varying both the weighting choices and the standard error estimation. In Table 6 I first reproduce the results from Table A9 in the published paper's supplementary materials. This table reports results for four different specifications – the base model, the interaction model, the pro-immigration only sample, and the anti-immigration only sample. Results from Table A9 are reproduced exactly.

Table 6: Spain Experiment: Regression Results With Weights and Classical (Non-Robust) Standard Errors (Computational Reproduction of Table A9)

	Base	Interaction model	Pro-immigration only	Anti-immigration only
Treatment	0.095***	0.211***	0.111***	0.103**
	(0.025)	(0.062)	(0.027)	(0.044)
Immigration	0.065***	0.075***		
	(0.005)	(0.007)		
Treatment x Immigration		–0.019**		
		(0.009)		
Intercept	0.238***	0.180***	0.784***	0.484***
	(0.033)	(0.044)	(0.019)	(0.030)
Num.Obs.	1196	1196	700	516
R2	0.154	0.157	0.023	0.011

* p < 0.1, ** p < 0.05, *** p < 0.01

In Table 7 I vary the weighting and standard error choices for the base model. I do the same for the interaction models in Table 8, for the pro-immigration sample in Table 9, and finally for the anti-immigration sample in Table 10. In each case I also reproduce the results from Table A9 as the first column ('replication').

Table 7: Spain Experiment: Base Model (Table A9, Column 1) Sensitivity

	Replication	No Weights	Weighted Robust SE	Unweighted Robust SE	Weighted Survey-Robust SE
Treatment	0.095***	0.026	0.095**	0.026	0.095**
	(0.025)	(0.023)	(0.042)	(0.023)	(0.042)
Immigration	0.065***	0.059***	0.065***	0.059***	0.065***
	(0.005)	(0.005)	(0.008)	(0.005)	(0.008)
Intercept	0.238***	0.366***	0.238***	0.366***	0.238***
	(0.033)	(0.034)	(0.057)	(0.039)	(0.057)
Num.Obs.	1196	1196	1196	1196	1196
R2	0.154	0.125	0.154	0.125	0.154
Weighted	Yes	No	Yes	No	Yes
HC3 Robust SEs	No	No	Yes	Yes	No
Survey Robust SEs	No	No	No	No	Yes

* p < 0.1, ** p < 0.05, *** p < 0.01

Table 8: Study 2 (Spain): Interaction Model (Table A9, Column 2) Sensitivity

	Replication	No Weights	Weighted Robust SE	Unweighted Robust SE	Weighted Survey-Robust SE
Treatment	0.211***	0.094	0.211*	0.094	0.211*
	(0.062)	(0.065)	(0.112)	(0.075)	(0.110)
Immigration	0.075***	0.065***	0.075***	0.065***	0.075***
	(0.007)	(0.007)	(0.011)	(0.007)	(0.011)
Treatment x Immigration	−0.019**	−0.010	−0.019	−0.010	−0.019
	(0.009)	(0.009)	(0.016)	(0.010)	(0.015)
Intercept	0.180***	0.331***	0.180**	0.331***	0.180**
	(0.044)	(0.047)	(0.071)	(0.053)	(0.070)
Num.Obs.	1196	1196	1196	1196	1196
R2	0.157	0.125	0.157	0.125	0.157
Weighted	Yes	No	Yes	No	Yes
HC3 Robust SEs	No	No	Yes	Yes	No
Survey Robust SEs	No	No	No	No	Yes

* p < 0.1, ** p < 0.05, *** p < 0.01

Table 9: Study 2 (Spain): Pro-Immigration Only Model (Table A9, Column 3) Sensitivity

	Replication	No Weights	Weighted Robust SE	Unweighted Robust SE	Weighted Survey-Robust SE
Treatment	0.111***	0.037	0.111**	0.037	0.111**
	(0.027)	(0.024)	(0.051)	(0.024)	(0.050)
Intercept	0.784***	0.871***	0.784***	0.871***	0.784***
	(0.019)	(0.017)	(0.041)	(0.018)	(0.040)
Num.Obs.	700	700	700	700	700
R2	0.023	0.004	0.023	0.004	0.023
Weighted	Yes	No	Yes	No	Yes
HC3 Robust SEs	No	No	Yes	Yes	No
Survey Robust SEs	No	No	No	No	Yes

* p < 0.1, ** p < 0.05, *** p < 0.01

Table 10: Study 2 (Spain): Anti-Immigration Only Model (Table A9, Column 4) Sensitivity

	Replication	No Weights	Weighted Robust SE	Unweighted Robust SE	Weighted Survey-Robust SE
Treatment	0.103**	0.034	0.103	0.034	0.103
	(0.044)	(0.043)	(0.069)	(0.043)	(0.068)
Intercept	0.484***	0.601***	0.484***	0.601***	0.484***
	(0.030)	(0.031)	(0.048)	(0.031)	(0.048)
Num.Obs.	516	516	516	516	516
R2	0.011	0.001	0.011	0.001	0.011
Weighted	Yes	No	Yes	No	Yes
HC3 Robust SEs	No	No	Yes	Yes	No
Survey Robust SEs	No	No	No	No	Yes

* p < 0.1, ** p < 0.05, *** p < 0.01

Re-analyses of Western Values

In Table 11 I re-analyse the results from Table A11 in the supplementary materials, which reports results for the ancillary and placebo outcomes. The only change here is to remove the weights,

given that the original version of Table A11 uses robust standard errors. As column 2 shows, the results for Western liberal values disappears once weights are removed.

Table 11: Study 2 (Spain): Regression Results Without Weights for Ancillary and Placebo Outcomes (Re-Estimation of Table A11)

	EU norms	Western liberal values	Green politics	Domestic violence protections	Spanish flag	Spanish military efforts
Treatment	0.508	0.493	−0.242	0.329	−0.245	0.432
	(0.471)	(0.426)	(0.455)	(0.472)	(0.594)	(0.595)
Immigration	0.256***	0.005	0.143***	0.210***	−0.343***	−0.083
	(0.048)	(0.045)	(0.049)	(0.050)	(0.060)	(0.062)
Treatment x Immigration	−0.088	−0.051	0.007	−0.061	0.009	−0.058
	(0.068)	(0.062)	(0.066)	(0.069)	(0.083)	(0.084)
Intercept	4.312***	6.902***	4.932***	4.915***	5.899***	5.011***
	(0.335)	(0.312)	(0.343)	(0.349)	(0.423)	(0.434)
Num.Obs.	1163	1171	1180	1179	1144	1113
R2	0.044	0.003	0.022	0.029	0.068	0.009
Weighted	No	No	No	No	No	No
HC3 Robust SEs	Yes	Yes	Yes	Yes	Yes	Yes
Survey Robust SEs	No	No	No	No	No	No

* p < 0.1, ** p < 0.05, *** p < 0.01

Interaction effects by weight bin

In Table 12 I present the results of the interaction model estimated on three sub-samples – those with low weights (under 0.01), those with high weights (over 3), and those mid weights (in-between).

Table 12: Study 2 (Spain): Heterogeneous Interaction Effects By Weight Bin

	Low Weights	Mid Weights	High Weights
Treatment	−0.019	−0.159	0.233**
	(0.103)	(0.336)	(0.118)
Immigration	0.050***	0.049*	0.075***
	(0.010)	(0.025)	(0.011)
Treatment x Immigration	0.001	0.017	−0.019
	(0.013)	(0.045)	(0.017)
Intercept	0.477***	0.487**	0.161**
	(0.077)	(0.206)	(0.074)
Num.Obs.	755	76	365
R2	0.096	0.122	0.164
Weighted	No	No	No
HC3 Robust SEs	Yes	Yes	Yes
Survey Robust SEs	No	No	No

* p < 0.1, ** p < 0.05, *** p < 0.01

Heterogeneous effects by age group

As noted in the main body of the comment, age is one of the primary factors that appears to differ between weight bins. A more detailed visualisation of the distribution of age across weight bins is presented in Figure 6. Essentially, every respondent who is over the age of 52 receives a high weight, those over the age of 35 but under 53 are much more likely to receive high weights than low weights, and those under 35 are much more likely to receive a low weight.

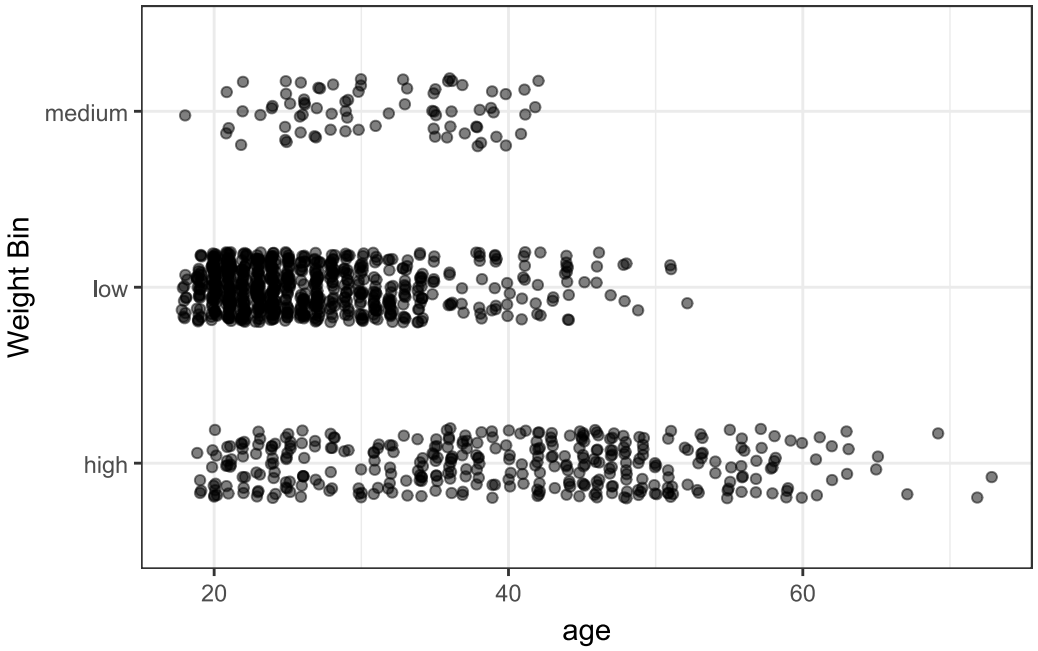

Figure 6: Study 2 (Spain): Age Distributions by Weight Bin

One might imagine that age is correlated with immigration sentiment, such that older people are more predisposed to be anti-immigrant than younger people. This is not the case in the study 2 data. As shown in **?@fig-imm1-by-age-es**, there is no meaningful association between `age` and `imm_1`, and in fact the oldest respondents are often the most positively disposed toward immigration.

A closer analysis is offered in Table 13, which shows the results of a simple linear regression of the outcome variable on the treatment indicator, subset into six age categories that were provided in the replication data. The treatment effect is only non-zero in the oldest age categories (45 years and older). The only statistically significant result is for the 45-54 category, and the point estimate is very large. While the point estimates for those in the two oldest categories are also quite large (albeit half the size of the 45-54 category), there are too few observations in this categories to say much with confidence. Notably it appears from this analysis that a small number of observations (around 150) likely drive the overall result presented in the paper. When these individuals are not up-weighted (and those who do not respond to the treatment are not down-weighted), the results unsurprisingly mostly attenuate.

Table 13: Study 2 (Spain): Heterogeneous Treatment Effects by Age Category

	Age < 25	Age 25-34	Age 35-44	Age 45-54	Age 55-64	Age >64
(Intercept)	0.859***	0.811***	0.673***	0.396***	0.682***	0.556**
	(0.025)	(0.027)	(0.047)	(0.063)	(0.095)	(0.162)
treat1	−0.039	0.004	−0.035	0.367***	0.131	0.194
	(0.034)	(0.039)	(0.071)	(0.089)	(0.147)	(0.214)
Num.Obs.	459	406	184	108	38	21
R2	0.003	0.000	0.001	0.139	0.021	0.042
R2 Adj.	0.001	−0.002	−0.004	0.131	−0.006	−0.009
AIC	388.9	393.7	253.6	143.6	50.7	33.1
BIC	401.3	405.7	263.2	151.6	55.6	36.2
Log.Lik.	−191.466	−193.858	−123.796	−68.795	−22.340	−13.558
F	1.287	0.012	0.241	17.084	0.790	0.826
RMSE	0.37	0.39	0.47	0.46	0.44	0.46

1. p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

It appears that the results in study 2 are largely on account of a combination of the weighting scheme targeting a small segment of the sample that appears to respond strongly to the stimulus. While this may be attributable to those individuals' characteristics (for example, that older people are theoretically more likely to respond to homonationalist appeals), it may well be down to pure chance.

Two things weigh against this deeper interpretation of these results. First, as was shown in **?@fig-imm1-by-age-es**, it is not the case that older voters are more likely to be more anti-immigration in the study 2 data. Second, for the UK (study 1) there is likewise no association between age and immigration sentiment, nor are there any coherent age heterogeneities. The relationship between age and immigration sentiment in the UK is shown in **?@fig-imm1-by-age-uk**, and the results of the binned regression analysis are shown in Table 14. In this context, the youngest age category has a statistically significant positive treatment effect, while the second youngest has a statistically significant negative effect. The remaining categories are not statistically significant, and the point estimates bounce around between 0 and 0.1, with no clear pattern.

Table 14: Study 1 (UK): Heterogeneous Treatment Effects by Age Category

	Age < 25	Age 25-34	Age 35-44	Age 45-54	Age 55-64	Age >64
(Intercept)	0.518***	0.797***	0.653***	0.699***	0.602***	0.597***
	(0.064)	(0.056)	(0.041)	(0.046)	(0.046)	(0.045)
treat1	0.189*	−0.149*	0.094	−0.038	0.073	−0.013
	(0.084)	(0.074)	(0.064)	(0.064)	(0.070)	(0.063)
Num.Obs.	131	152	215	212	194	244
R2	0.037	0.026	0.010	0.002	0.006	0.000
R2 Adj.	0.030	0.020	0.005	−0.003	0.000	−0.004
AIC	182.6	192.9	281.2	284.2	272.1	352.1
BIC	191.2	202.0	291.3	294.3	281.9	362.6
Log.Lik.	−88.297	−93.459	−137.601	−139.107	−133.027	−173.040
F	4.993	4.062	2.183	0.357	1.081	0.040
RMSE	0.47	0.45	0.46	0.47	0.48	0.49

1. p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Corrected Figure 6 (No Weights)

In Figure 7 I present the corrected (linear regression, robust standard errors) version of Figure 6 from the published paper, but with weights removed.

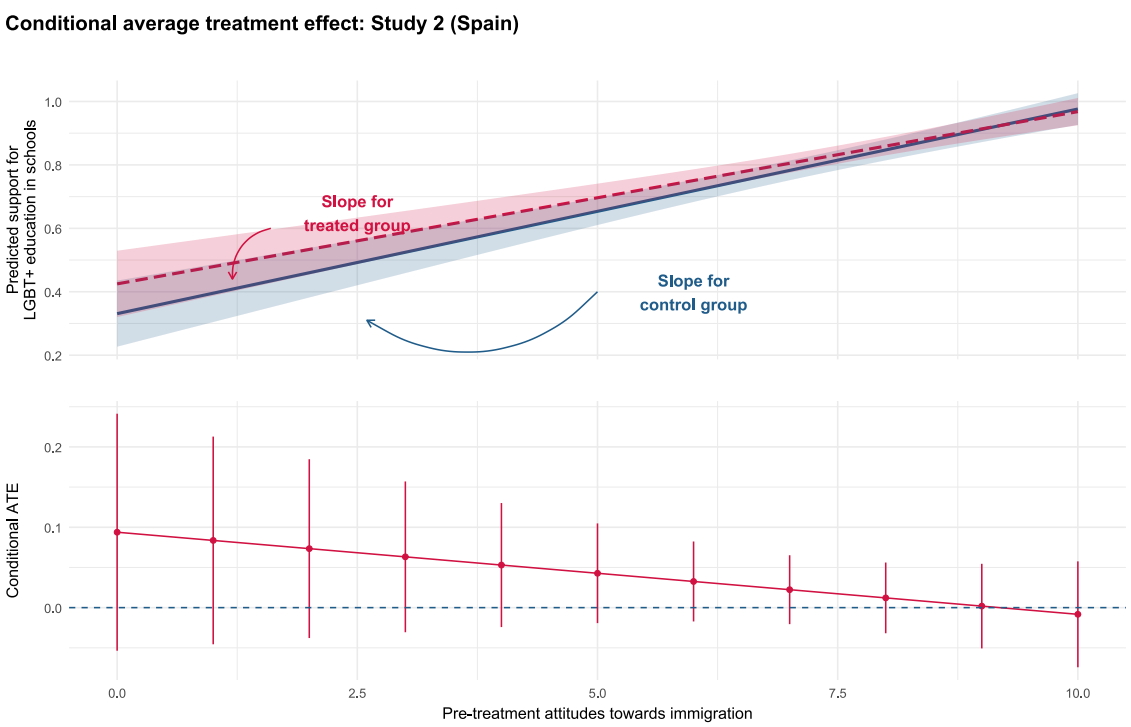

Figure 7: Study 2 (Spain): Corrected Reproduction of Figure 6 Without Weighting

Appendix 3: All Code Used In This Comment

This comment was produced as a dynamic document using Quarto. This appendix includes all code run.

```
# load libraries
library(tidyverse);   library(jtools);   library(ggpubr);   library(ggrepel);
library(patchwork); library(gt); library(modelsummary); library(interactions);
library(margins); library(skimr); library(survey); library(estimatr)

# modelsummary options
options(modelsummary_model_labels = "Computer Modern")

# set seed exactly as per replication materials
set.seed(1)
# load both datasets
uk <- read_csv("study1_data.csv")
load("study2_data.Rda")

# set color palette as per replication materials
colors<- c("#205C8A", "#d11141")

# cleaning per replication materials
uk <- uk%>%
  mutate(treat= as.factor(treatment),
         treatnum= as.numeric(treatment),
         gender= as.factor(gender),
         degree= as.factor(degree),
         nonwhite= as.factor(nonwhite),
         queer= as.factor(queer),
         relig= as.factor(relig),
         religion= as.factor(religion),
         race= as.factor(race),
         fourarm= as.factor(fourarm),
         immbelow= as.factor(immbelow),
         imm3= as.factor(imm3),
         region= as.factor(region),
         voterecall= as.factor(voterecall),
         brexit= as.factor(brexit),
         ideology= as.factor(ideology),
         agecat= as.factor(agecat))

# cleaning per replication materials
spain <- spain%>%
  mutate(treat= as.factor(treat),
         treatnum= as.numeric(treat),
         gender= as.factor(gender),
         supportcat= as.factor(support),
         agecat= as.factor(agecat),
```

```
         child= as.factor(child),
         immdum= as.factor(immdum),
         imm5= as.factor(imm5),
         imm3= as.factor(imm3),
         foreignborn= as.factor(foreignborn),
         CCAA= as.factor(CCAA),
         queer= as.factor(queer))
# subset uk data per replication materials
treat <- subset(uk, treatnum==1)
control <- subset(uk, treatnum==0)
proimm <- subset(uk, immbelow==0)
noproimm <- subset(uk, immbelow==1)

# analyse uk data per replication materials
modelsub1<- lm(support ~ treat, data=proimm)
proimm$predictedb <- predict(modelsub1, proimm)

modelsub2<- lm(support ~ treat, data=noproimm)
noproimm$predictedb <- predict(modelsub2, noproimm)

# subset again per replication materials
treatsub1 <- subset(proimm, treatnum==1)
controlsub1 <- subset(proimm, treatnum==0)
treatsub2 <- subset(noproimm, treatnum==1)
controlsub2 <- subset(noproimm, treatnum==0)
# subset spain data per replication materials
treatES <- subset(spain, treat==1)
controlES <- subset(spain, treat==0)
proimmES <- subset(spain, immdum==1)
noproimmES <- subset(spain, immdum==0)

# analyse spain data, but use lm() not glm() so that Spain and UK analyses are
exactly the same:
modelsub1ES <- lm(support ~ treat, weight=nationalweight, data=proimmES)
proimmES$predictedb <- predict(modelsub1ES, proimmES)

modelsub2ES <- lm(support ~ treat, weight=nationalweight, data=noproimmES)
noproimmES$predictedb <- predict(modelsub2ES, noproimmES)

# subset again per replication materials
treatsub1ES <- subset(proimmES, treat==1)
controlsub1ES <- subset(proimmES, treat==0)
treatsub2ES <- subset(noproimmES, treat==1)
controlsub2ES <- subset(noproimmES, treat==0)
# create weights plot
weights_plot <- ggplot(spain, aes(x=nationalweight)) +
  geom_histogram(bins=30, fill="black", color="black") +
  theme_minimal() +
```

```
  labs(x = "Weight",
       y = "Frequency")

weights_plot
df <- data.frame(
  Study = c("Study 1:", rep("",6),
            "Study 2:", rep("",7)
            ),
  Item = c(
  "", "Figure 3", "Figure 4", "Figure A4", "Table A7", "Table A8",
  "",
  "", "Figure 6", "Figure 7", "Figure 8", "Figure A5", "Table A9", "Table A10",
"Table A11"
  ),
  RobustSEs = c("", "No", "Yes", "No", "No", "Yes",
                "",
                "", "No", "No", "No", "No", "No", "No", "Yes"),
  Notes = c(
    "", #study 1:
    "90% CIs shown",
    "Text overlay uses classical SE, 90% CIs shown",
    "90% CIs shown",
    "", "",
    "",
    "", #study 2:
    "90% CIs shown",
    "Text overlay uses classical SE, 90% CIs shown",
    "90% CIs shown",
    "90% CIs shown",
    "", "", ""
  ),
  stringsAsFactors = FALSE
)

# Create the gt table
df %>%
  gt() %>%
  cols_align(
    align = "left",
    columns = c(Study, Item, Notes)
  ) %>%
  cols_label(
    Study = "",
    Item = "Figure/Table",
    RobustSEs = "Robust SEs?",
    Notes = "Notes"
  )
# computational reproduction of Table A9:
```

```r
models_rep <- list(
  'Base' = lm(support ~ treat + imm_1, data=spain, weight = nationalweight),
    'Interaction model' = lm(support ~ treat*imm_1, data=spain, weight =
nationalweight),
    'Pro-immigration only' = lm(support ~ treat, data=proimmES, weight =
nationalweight),
   'Anti-immigration only' = lm(support ~ treat, data=noproimmES, weight =
nationalweight)
)

# reproduction by analysis, with varying researcher choices:
models_base <- list(
    'Replication' = lm(support ~ treat + imm_1, data=spain, weight =
nationalweight),
  'No Weights' = lm(support ~ treat + imm_1, data=spain),
  'Weighted Robust SE' = estimatr::lm_robust(support ~ treat + imm_1, data=spain,
weight = nationalweight, se_type = "HC3"),
    'Unweighted Robust SE' = estimatr::lm_robust(support ~ treat + imm_1,
data=spain, se_type = "HC3"),
    'Weighted Survey-Robust SE' = survey::svyglm(support ~ treat + imm_1,
design=svydesign(ids=~1, weights=~nationalweight, data=spain))
)

models_int <- list(
  'Replication' = lm(support ~ treat*imm_1, data=spain, weight = nationalweight),
  'No Weights' = lm(support ~ treat*imm_1, data=spain),
  'Weighted Robust SE' = estimatr::lm_robust(support ~ treat*imm_1, data=spain,
weight = nationalweight, se_type = "HC3"),
  'Unweighted Robust SE' = estimatr::lm_robust(support ~ treat*imm_1, data=spain,
se_type = "HC3"),
    'Weighted Survey-Robust SE' = survey::svyglm(support ~ treat*imm_1,
design=svydesign(ids=~1, weights=~nationalweight, data=spain))
)

models_proimmes <- list(
  'Replication' = lm(support ~ treat, data=proimmES, weight = nationalweight),
  'No Weights' = lm(support ~ treat, data=proimmES),
   'Weighted Robust SE' = estimatr::lm_robust(support ~ treat, data=proimmES,
weight = nationalweight, se_type = "HC3"),
  'Unweighted Robust SE' = estimatr::lm_robust(support ~ treat, data=proimmES,
se_type = "HC3"),
      'Weighted Survey-Robust SE' = survey::svyglm(support ~ treat,
design=svydesign(ids=~1, weights=~nationalweight, data=proimmES))
)

models_noproimmes <- list(
  'Replication' = lm(support ~ treat, data=noproimmES, weight = nationalweight),
  'No Weights' = lm(support ~ treat, data=noproimmES),
```

```
  'Weighted Robust SE' = estimatr::lm_robust(support ~ treat, data=noproimmES,
weight = nationalweight, se_type = "HC3"),
  'Unweighted Robust SE' = estimatr::lm_robust(support ~ treat, data=noproimmES,
se_type = "HC3"),
      'Weighted Survey-Robust SE' = survey::svyglm(support ~ treat,
design=svydesign(ids=~1, weights=~nationalweight, data=noproimmES))
)

# ancillary mechanism test (A11):
mech <- list(
  'EU norms' = estimatr::lm_robust(pride_valoresUE ~ treat*imm_1, data=spain,
se_type = "HC3"),
 'Western liberal values' = estimatr::lm_robust(pride_libertadOCC ~ treat*imm_1,
data=spain, se_type = "HC3"),
  'Green politics' = estimatr::lm_robust(pride_verde ~ treat*imm_1, data=spain,
se_type = "HC3"),
    'Domestic violence protections' = estimatr::lm_robust(pride_viomach ~
treat*imm_1, data=spain, se_type = "HC3"),
  'Spanish flag' = estimatr::lm_robust(pride_bandera ~ treat*imm_1, data=spain,
se_type = "HC3"),
  'Spanish military efforts' = estimatr::lm_robust(pride_mili ~ treat*imm_1,
data=spain, se_type = "HC3")
)

get_stars <- function(est, se) {
  t <- abs(est / se)
  stars <- if (t > 2.576) "***"
           else if (t > 1.96) "**"
           else if (t > 1.645) "*"
           else ""
  c(paste0(round(est, 3), stars), paste0("(", round(se, 3), ")"))
}

get_stars_interaction <- function(est1, se1, est2, se2) {
  stars1 <- if (abs(est1/se1) > 2.576) "***"
            else if (abs(est1/se1) > 1.96) "**"
            else if (abs(est1/se1) > 1.645) "*"
            else ""
  stars2 <- if (abs(est2/se2) > 2.576) "***"
            else if (abs(est2/se2) > 1.96) "**"
            else if (abs(est2/se2) > 1.645) "*"
            else ""
  est_str <- paste0(round(est1, 3), stars1, " | ", round(est2, 3), stars2)
  se_str <- paste0("(", round(se1, 3), ") | (", round(se2, 3), ")")
  c(est_str, se_str)
}

table_data <- data.frame(
```

```
  Choices = c(
    "Weighted + Classical SE",
   "",
    "Unweighted + Classical SE",
   "",
    "Weighted + Robust SE",
   "",
    "Unweighted + Robust SE",
   "",
    "Weighted + Survey-R SE",
   ""
  ),

  Baseline = c(
                  get_stars(summary(models_base$Replication)$coefficients[2,1],
summary(models_base$Replication)$coefficients[2,2]),
              get_stars(summary(models_base$`No   Weights`)$coefficients[2,1],
summary(models_base$`No Weights`)$coefficients[2,2]),
       get_stars(summary(models_base$`Weighted  Robust  SE`)$coefficients[2,1],
summary(models_base$`Weighted Robust SE`)$coefficients[2,2]),
      get_stars(summary(models_base$`Unweighted  Robust  SE`)$coefficients[2,1],
summary(models_base$`Unweighted Robust SE`)$coefficients[2,2]),
                          get_stars(summary(models_base$`Weighted      Survey-
Robust   SE`)$coefficients[2,1],   summary(models_base$`Weighted   Survey-Robust
SE`)$coefficients[2,2])
  ),

  Interaction = c(
    get_stars_interaction(
                             summary(models_int$Replication)$coefficients[2,1],
summary(models_int$Replication)$coefficients[2,2],
                             summary(models_int$Replication)$coefficients[4,1],
summary(models_int$Replication)$coefficients[4,2]),
    get_stars_interaction(
     summary(models_int$`No Weights`)$coefficients[2,1], summary(models_int$`No
Weights`)$coefficients[2,2],
     summary(models_int$`No Weights`)$coefficients[4,1], summary(models_int$`No
Weights`)$coefficients[4,2]),
    get_stars_interaction(
                summary(models_int$`Weighted   Robust   SE`)$coefficients[2,1],
summary(models_int$`Weighted Robust SE`)$coefficients[2,2],
                summary(models_int$`Weighted   Robust   SE`)$coefficients[4,1],
summary(models_int$`Weighted Robust SE`)$coefficients[4,2]),
    get_stars_interaction(
               summary(models_int$`Unweighted   Robust   SE`)$coefficients[2,1],
summary(models_int$`Unweighted Robust SE`)$coefficients[2,2],
               summary(models_int$`Unweighted   Robust   SE`)$coefficients[4,1],
summary(models_int$`Unweighted Robust SE`)$coefficients[4,2]),
```

```
    get_stars_interaction(
          summary(models_int$`Weighted  Survey-Robust  SE`)$coefficients[2,1],
summary(models_int$`Weighted Survey-Robust SE`)$coefficients[2,2],
          summary(models_int$`Weighted  Survey-Robust  SE`)$coefficients[4,1],
summary(models_int$`Weighted Survey-Robust SE`)$coefficients[4,2])
  ),

  ProImmigration = c(
             get_stars(summary(models_proimmes$Replication)$coefficients[2,1],
summary(models_proimmes$Replication)$coefficients[2,2]),
          get_stars(summary(models_proimmes$`No   Weights`)$coefficients[2,1],
summary(models_proimmes$`No Weights`)$coefficients[2,2]),
    get_stars(summary(models_proimmes$`Weighted Robust SE`)$coefficients[2,1],
summary(models_proimmes$`Weighted Robust SE`)$coefficients[2,2]),
   get_stars(summary(models_proimmes$`Unweighted Robust SE`)$coefficients[2,1],
summary(models_proimmes$`Unweighted Robust SE`)$coefficients[2,2]),
                  get_stars(summary(models_proimmes$`Weighted    Survey-Robust
SE`)$coefficients[2,1],    summary(models_proimmes$`Weighted    Survey-Robust
SE`)$coefficients[2,2])
  ),

  AntiImmigration = c(
           get_stars(summary(models_noproimmes$Replication)$coefficients[2,1],
summary(models_noproimmes$Replication)$coefficients[2,2]),
         get_stars(summary(models_noproimmes$`No  Weights`)$coefficients[2,1],
summary(models_noproimmes$`No Weights`)$coefficients[2,2]),
   get_stars(summary(models_noproimmes$`Weighted Robust SE`)$coefficients[2,1],
summary(models_noproimmes$`Weighted Robust SE`)$coefficients[2,2]),
                    get_stars(summary(models_noproimmes$`Unweighted    Robust
SE`)$coefficients[2,1],      summary(models_noproimmes$`Unweighted      Robust
SE`)$coefficients[2,2]),
                get_stars(summary(models_noproimmes$`Weighted    Survey-Robust
SE`)$coefficients[2,1],    summary(models_noproimmes$`Weighted    Survey-Robust
SE`)$coefficients[2,2])
  )
)


legend_row <- rep("", ncol(table_data))
legend_row[1] <- "* p < 0.1, ** p < 0.05, *** p < 0.01. SEs in parentheses."

# Bind it to the bottom of your existing table
table_data_final <- rbind(table_data, legend_row)

# If your table columns are factors, convert them to character first to avoid
issues
table_data_final[] <- lapply(table_data_final, as.character)
```

```r
# Create the table
table_data_final |>
  gt() |>
  cols_align(
    align = "center",
    columns = c(Baseline, Interaction, ProImmigration, AntiImmigration)
  )  %>%
  cols_label(
    Choices = "Researcher Choices",
    Baseline = "Base Model",
    Interaction = "Interaction",
    ProImmigration = "Pro-immig.",
    AntiImmigration = "Anti-immig.",
  ) %>%
 tab_style(
  style = cell_borders(
    sides = "top",
    color = "black",
    weight = px(1.5)
  ),
  locations = cells_body(
    rows = nrow(table_data_final)  # target legend row
  )
 )

# create bins of weights:
spain <- spain %>%
  mutate(weightbins = case_when(
    nationalweight >= 0 & nationalweight < 0.01 ~ "low",
    nationalweight >= 0.1 & nationalweight < 3 ~ "medium",
    nationalweight >= 3 ~ "high",
    TRUE ~ "other"
  ))

# re-subset
proimmES <- subset(spain, immdum==1)
noproimmES <- subset(spain, immdum==0)

# build tables - anti-immigration subsample
models_hetfx_anti <- list(
         'Low    Weights'    =    estimatr::lm_robust(support    ~    treat,
data=noproimmES[noproimmES$weightbins=="low",], se_type = "HC3"),
         'Mid    Weights'    =    estimatr::lm_robust(support    ~    treat,
data=noproimmES[noproimmES$weightbins=="medium",], se_type = "HC3"),
         'High    Weights'    =    estimatr::lm_robust(support    ~    treat,
data=noproimmES[noproimmES$weightbins=="high",], se_type = "HC3")
)
```

```r
# build tables - pro-immigration subsample
models_hetfx_pro <- list(
         'Low    Weights'     =     estimatr::lm_robust(support     ~     treat,
data=proimmES[proimmES$weightbins=="low",], se_type = "HC3"),
         'Mid    Weights'     =     estimatr::lm_robust(support     ~     treat,
data=proimmES[proimmES$weightbins=="medium",], se_type = "HC3"),
         'High    Weights'     =     estimatr::lm_robust(support     ~     treat,
data=proimmES[proimmES$weightbins=="high",], se_type = "HC3")
)


# build tables - interaction hetfx
models_hetfx_int <- list(
       'Low    Weights'    =    estimatr::lm_robust(support    ~    treat*imm_1,
data=spain[spain$weightbins=="low",], se_type = "HC3"),
       'Mid    Weights'    =    estimatr::lm_robust(support    ~    treat*imm_1,
data=spain[spain$weightbins=="medium",], se_type = "HC3"),
       'High    Weights'    =    estimatr::lm_robust(support    ~    treat*imm_1,
data=spain[spain$weightbins=="high",], se_type = "HC3")
)
suppressWarnings(
  modelsummary(models_hetfx_anti, output = "gt",
             coef_map = c('treat1' = 'Treatment',
                          '(Intercept)' = 'Intercept'),
             gof_omit = "BIC|AIC|R2 Adj.|F|RMSE|Log.Lik.",
             stars = c('*'=.1, "**"=.05, "***"=.01),
             add_rows = tribble(~term,  ~Base,  ~Base, ~Base,
               "Weighted", "No", "No", "No",
               "HC3 Robust SEs", "Yes", "Yes", "Yes",
               "Survey Robust SEs", "No", "No", "No",
             ))
)
suppressWarnings(
  modelsummary(models_hetfx_pro, output = "gt",
             coef_map = c('treat1' = 'Treatment',
                          '(Intercept)' = 'Intercept'),
             gof_omit = "BIC|AIC|R2 Adj.|F|RMSE|Log.Lik.",
             stars = c('*'=.1, "**"=.05, "***"=.01),
             add_rows = tribble(~term,  ~Base,  ~Base, ~Base,
               "Weighted", "No", "No", "No",
               "HC3 Robust SEs", "Yes", "Yes", "Yes",
               "Survey Robust SEs", "No", "No", "No",
             ))
)
# create a balance table where we show the mean of imm_1, age, edu, gender, and
other covariates by weightbin:
bal_table <- spain %>%
```

```
  mutate(gender = as.numeric(as.character(gender)) - 1) %>%
  group_by(weightbins) %>%
  summarise_at(
    c("age", "gender", "edu", "child", "foreignborn", "queer", "imm_1"),
    ~round(mean(as.numeric(as.character(.)), na.rm = T),2)
  ) %>%
# transpose the table so the weightbins are the columns, and the covariates the
rows:
  pivot_longer(cols = c(age, gender, edu, child, foreignborn, queer, imm_1)) %>%
  pivot_wider(names_from = weightbins, values_from = value) %>%
  relocate(low, .before = high) %>%
  rename(`Low Weight Bin` = low, `Medium Weight Bin` = medium, `High Weight Bin`
= high, `Variable` = name)
gt(bal_table)
# correct the code to use a linear regression lm() and not logistic regression.
model1 <- lm(support ~ treat*imm_1, data=uk)
# use lm_robust for robust SEs that can be used by margins(). Set se_type =
"HC3" to be consistent with summ(.,robust=TRUE). Point estimates are of course
numerically identical bar rounding.
model1_robust <- estimatr::lm_robust(support ~ treat*imm_1, data=uk, se_type =
"HC3")

# create plots per the replication materials, adding in the CI and making SE
robust. Must use lm class object here not lm_robust object.
pred<- interact_plot(model1, pred = imm_1, modx = treat, interval = TRUE, robust
= TRUE,
                     colors = colors) +
  labs(title="",
       y="Predicted support for\nLGBT+ education in schools",
       x="")+
  theme_minimal()+
  theme(legend.position = "none",
        axis.text.x =element_blank())+
  annotate(
    geom="text", x = 2.5, y = .65, size = 4, color = "#d11141", fontface=2,
    label = "Slope for\ntreated group")+
  annotate(
    geom = "curve", x =1.6, y = .6, xend = 1.2, yend = .44,
    curvature = .4, arrow = arrow(length = unit(2, "mm")), colour="#d11141")+
  annotate(
    geom="text", x = 6, y = .4, size = 4, color = "#205C8A", fontface=2,
    label = "Slope for\ncontrol group")+
  annotate(
    geom = "curve", x =5, y = .4, xend = 2.52, yend =.38,
    curvature = -.4, arrow = arrow(length = unit(2, "mm")), colour="#205C8A")

gg_df <-
  # update to robust object
```

```
  model1_robust %>%
  margins(at = list(imm_1 = seq(0, 10, by = 1))) %>%
  summary %>%
  as.data.frame() %>%
  filter(factor == "treat1")

ame<- ggplot(gg_df, aes(imm_1, AME)) +
  geom_point(colour="#d11141") +
  geom_line(colour="#d11141") +
  coord_cartesian(xlim = c(0, 10), ylim = c(-.25, .25)) +
  # change to 95% ci:
   geom_errorbar(aes(ymax = (AME-SE*1.96), ymin = (AME+SE*1.96)), width = 0,
colour="#d11141") +
  geom_hline(yintercept = 0, linetype = "dashed", colour="#205C8A") +
  xlab("Pre-treatment attitudes towards immigration")+
  ylab("Conditional ATE") +
  theme_minimal()
pred/ame+
  plot_annotation(title = 'Conditional average treatment effect: Study 1 (UK)',
                  theme = theme(plot.title = element_text(size = 14, face="bold")))
# correct the code to use a linear regression lm() and not logistic regression.
modelES <- lm(support ~ treat*imm_1, data=spain, weight=nationalweight)
# use lm_robust for robust SEs that can be used by margins(). Set se_type =
"HC3" to be consistent with summ(.,robust=TRUE). Point estimates are of course
numerically identical bar rounding.
modelES_robust  <-  estimatr::lm_robust(support  ~  treat*imm_1,  data=spain,
weight=nationalweight, se_type = "HC3")

# create plots per the replication materials, adding in the CI and making SE
robust:
predES<- interact_plot(modelES, pred = imm_1, modx = treat, interval = TRUE,
robust = TRUE,
                       colors = colors)+
  labs(title="",
       y="Predicted support for\nLGBT+ education in schools",
       x="")+
  theme_minimal()+
  theme(legend.position = "none",
        axis.text.x =element_blank())+
  annotate(
    geom="text", x = 2.5, y = .65, size = 4, color = "#d11141", fontface=2,
    label = "Slope for\ntreated group")+
  annotate(
    geom = "curve", x =1.6, y = .6, xend = 1.2, yend = .44,
    curvature = .4, arrow = arrow(length = unit(2, "mm")), colour="#d11141")+
  annotate(
    geom="text", x = 6, y = .4, size = 4, color = "#205C8A", fontface=2,
    label = "Slope for\ncontrol group")+
```

```
  annotate(
    geom = "curve", x =5, y = .4, xend = 2.6, yend =.31,
    curvature = -.4, arrow = arrow(length = unit(2, "mm")), colour="#205C8A")

gg_df <-
  # update to robust object
  modelES_robust %>%
  margins(at = list(imm_1 = seq(0, 10, by = 1))) %>%
  summary %>%
  as.data.frame() %>%
  filter(factor == "treat1")

ameES<- ggplot(gg_df, aes(imm_1, AME)) +
  geom_point(colour="#d11141") +
  geom_line(colour="#d11141") +
  coord_cartesian(xlim = c(0, 10)) +
  # change to 95% ci:
   geom_errorbar(aes(ymax = (AME-SE*1.96), ymin = (AME+SE*1.96)), width = 0,
colour="#d11141") +
  geom_hline(yintercept = 0, linetype = "dashed", colour="#205C8A") +
  xlab("Pre-treatment attitudes towards immigration")+
  ylab("Conditional ATE") +
  theme_minimal()
predES/ameES+
   plot_annotation(title = 'Conditional average treatment effect: Study 2
(Spain)',
                theme = theme(plot.title = element_text(size = 14, face="bold")))
proimmplot<- effect_plot(model = modelsub1, pred = treat, robust=TRUE,
                         cat.geom="point", cat.interval.geom="linerange",
                         # correct the 95% ci:
                        colors="black", cat.pred.point.size=3, int.width = .95)+
  labs(title = paste0("Pro-immigrant voters (N = ", nrow(proimm),")"))+
  ylab("Support for LGBT+ education in schools (0-1)")+
  xlab("")+
  scale_y_continuous(limits=c(-0.1,1.1), breaks=c(0, .2, .4, .6, .8, 1)) +
  scale_x_discrete(labels=c("0" = "Control", "1" = "Treatment"))+
  # correct the y variable:
  geom_jitter(data=treatsub1, aes(x=treat, y=support),
              height=.05, width=.2, alpha=.35, shape=20,
              pch=21, size=2, color="#d11141")+
  # correct the y variable:
  geom_jitter(data=controlsub1, aes(x=treat, y=support),
              height=.05, width=.2, alpha=.35, shape=20,
              pch=21, size=2, color="#205C8A")+
  geom_bracket(xmin = c("0"), xmax = c("1"),
               y.position = c(.45), label = c("ATE=-.06*"),
               tip.length =-0.05,
               color="black")+
```

```
  theme_minimal()+
  theme(axis.title.y = element_text(face="bold"),
        axis.text.x = element_text(face="bold"))

noproimmplot<- effect_plot(model = modelsub2, pred = treat, robust=TRUE,
                           cat.geom="point", cat.interval.geom="linerange",
                           # correct the 95% ci:
                              colors="black", cat.pred.point.size=3, int.width
= .95)+
  labs(title = paste0("Anti-immigrant voters (N = ", nrow(noproimm),")"))+
  ylab("Support for LGBT+ education in schools (0-1)")+
  xlab("")+
  scale_y_continuous(limits=c(-0.1,1.1), breaks=c(0, .2, .4, .6, .8, 1)) +
  scale_x_discrete(labels=c("0" = "Control", "1" = "Treatment"))+
  # correct the y variable:
  geom_jitter(data=treatsub2, aes(x=treat, y=support),
              height=.05, width=.2, alpha=.35, shape=20,
              pch=21, size=2, color="#d11141")+
  # correct the y variable:
  geom_jitter(data=controlsub2, aes(x=treat, y=support),
              height=.05, width=.2, alpha=.35, shape=20,
              pch=21, size=2, color="#205C8A")+
  geom_bracket(xmin = c("0"), xmax = c("1"),
               y.position = c(.79), label = c("ATE=.10**"),
               tip.length =0.05,
               color="black")+
  theme_minimal()+
  theme(axis.title.y = element_blank(),
        axis.text.y = element_blank(),
        axis.text.x = element_text(face="bold"))

proimmplot+noproimmplot+
  plot_annotation(title = 'Effect of out-group treatment among:',
                  caption = "Lines are 95% confidence intervals.Treatment group
outcome statistically distinct at p<.1(*), p<0.05(**), & p<0.01(***)",
                theme = theme(plot.title = element_text(size = 14, face="bold")))

proimmplotES<- effect_plot(model = modelsub1ES, pred = treat,
                           cat.geom="point", cat.interval.geom="linerange",
                           # correct 95% ci and make them robust:
                              colors="black", cat.pred.point.size=2, int.width
= .95, robust = TRUE)+
  labs(title = paste0("Pro-immigrant voters (N = ", nrow(proimmES),")"))+
  ylab("Support for LGBT+ education in schools (0-1)")+
  xlab("")+
  scale_y_continuous(limits=c(-0.1,1.1), breaks=c(0, .2, .4, .6, .8, 1)) +
  scale_x_discrete(labels=c("0" = "Control", "1" = "Treatment"))+
  # correct the y variable:
```

```
  geom_jitter(data=treatsub1ES, aes(x=treat, y=support),
              height=.05, width=.2, alpha=.35, shape=20,
              pch=21, size=2,  color="#d11141")+
  # correct the y variable:
  geom_jitter(data=controlsub1ES, aes(x=treat, y=support),
              height=.05, width=.2, alpha=.35, shape=20,
              pch=21, size=2, color="#205C8A")+
  geom_bracket(xmin = c("0"), xmax = c("1"),
               # correct significance from *** to ** due to robust SEs:
               y.position = c(.45), label = c("ATE=.11**"),
               tip.length =-0.05,
               color="black")+
  theme_minimal() +
  theme(axis.title.y = element_text(face="bold"),
        axis.text.x = element_text(face="bold"))

noproimmplotES<- effect_plot(model = modelsub2ES, pred = treat,
                             cat.geom="point", cat.interval.geom="linerange",
                            # correct 95% ci and make them robust:
                              colors="black", cat.pred.point.size=2, int.width
= .95, robust = TRUE)+
  labs(title = paste0("Anti-immigrant voters (N = ", nrow(noproimmES),")"))+
  ylab("Support for LGBT+ education in schools (0-1)")+
  xlab("")+
  scale_y_continuous(limits=c(-0.1,1.1), breaks=c(0, .2, .4, .6, .8, 1)) +
  scale_x_discrete(labels=c("0" = "Control", "1" = "Treatment"))+
  # correct the y variable:
  geom_jitter(data=treatsub2ES, aes(x=treat, y=support),
              height=.05, width=.2, alpha=.35, shape=20,
              pch=21, size=2, color="#d11141")+
  # correct the y variable:
  geom_jitter(data=controlsub2ES, aes(x=treat, y=support),
              height=.05, width=.2, alpha=.35, shape=20,
              pch=21, size=2, color="#205C8A")+
  geom_bracket(xmin = c("0"), xmax = c("1"),
               # correct significance from *** to [] due to robust SEs:
               y.position = c(.79), label = c("ATE=.10"),
               tip.length = 0.05,
               color="black")+
  theme_minimal()+
  theme(axis.title.y = element_blank(),
        axis.text.y = element_blank(),
        axis.text.x = element_text(face="bold"))
proimmplotES+noproimmplotES+
  plot_annotation(title = 'Effect of out-group treatment among:',
                  caption="Lines are 95% confidence intervals. Treatment group
outcome statistically distinct at p<.1(*), p<0.05(**), & p<0.01(***)",
              theme = theme(plot.title = element_text(size = 14, face="bold")))
```

```
modelsummary(models_rep, output = "gt",
               coef_map = c('treat1' = 'Treatment', 'imm_1' = 'Immigration',
'treat1:imm_1' = 'Treatment x Immigration', '(Intercept)' = 'Intercept'),
             gof_omit = "BIC|AIC|R2 Adj.|F|RMSE|Log.Lik.",
             stars = c('*'=.1, "**"=.05, "***"=.01))
suppressWarnings(
  modelsummary(models_base, output = "gt",
               coef_map = c('treat1' = 'Treatment', 'imm_1' = 'Immigration',
'(Intercept)' = 'Intercept'),
             gof_omit = "BIC|AIC|R2 Adj.|F|RMSE|Log.Lik.",
             stars = c('*'=.1, "**"=.05, "***"=.01),
             add_rows = tribble(~term,  ~Base,  ~Base, ~Base, ~Base, ~Base,
               "Weighted", "Yes", "No", "Yes", "No", "Yes",
               "HC3 Robust SEs", "No", "No", "Yes", "Yes", "No",
               "Survey Robust SEs", "No", "No", "No", "No", "Yes"
             ))
)

suppressWarnings(
  modelsummary(models_int, output = "gt",
               coef_map = c('treat1' = 'Treatment', 'imm_1' = 'Immigration',
'treat1:imm_1' = 'Treatment x Immigration',
                            '(Intercept)' = 'Intercept'),
             gof_omit = "BIC|AIC|R2 Adj.|F|RMSE|Log.Lik.",
             stars = c('*'=.1, "**"=.05, "***"=.01),
             add_rows = tribble(~term,  ~Base,  ~Base, ~Base, ~Base, ~Base,
               "Weighted", "Yes", "No", "Yes", "No", "Yes",
               "HC3 Robust SEs", "No", "No", "Yes", "Yes", "No",
               "Survey Robust SEs", "No", "No", "No", "No", "Yes"
             ))
)

suppressWarnings(
  modelsummary(models_proimmes, output = "gt",
             coef_map = c('treat1' = 'Treatment', 'imm_1' = 'Immigration',
                          '(Intercept)' = 'Intercept'),
             gof_omit = "BIC|AIC|R2 Adj.|F|RMSE|Log.Lik.",
             stars = c('*'=.1, "**"=.05, "***"=.01),
             add_rows = tribble(~term,  ~Base,  ~Base, ~Base, ~Base, ~Base,
               "Weighted", "Yes", "No", "Yes", "No", "Yes",
               "HC3 Robust SEs", "No", "No", "Yes", "Yes", "No",
               "Survey Robust SEs", "No", "No", "No", "No", "Yes"
             ))
)

suppressWarnings(
  modelsummary(models_noproimmes, output = "gt",
```

```
             coef_map = c('treat1' = 'Treatment', 'imm_1' = 'Immigration',
                          '(Intercept)' = 'Intercept'),
             gof_omit = "BIC|AIC|R2 Adj.|F|RMSE|Log.Lik.",
             stars = c('*'=.1, "**"=.05, "***"=.01),
             add_rows = tribble(~term,  ~Base,  ~Base, ~Base, ~Base, ~Base,
               "Weighted", "Yes", "No", "Yes", "No", "Yes",
               "HC3 Robust SEs", "No", "No", "Yes", "Yes", "No",
               "Survey Robust SEs", "No", "No", "No", "No", "Yes"
            ))
)

suppressWarnings(
modelsummary(mech, output = "gt",
                coef_map = c('treat1' = 'Treatment', 'imm_1' = 'Immigration',
'treat1:imm_1' = 'Treatment x Immigration',
                          '(Intercept)' = 'Intercept'),
             gof_omit = "BIC|AIC|R2 Adj.|F|RMSE|Log.Lik.",
             stars = c('*'=.1, "**"=.05, "***"=.01),
               add_rows = tribble(~term,  ~Base,  ~Base, ~Base, ~Base, ~Base,
~Base,
               "Weighted", "No", "No", "No", "No", "No", "No",
               "HC3 Robust SEs", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes",
               "Survey Robust SEs", "No", "No", "No", "No", "No", "No"
             ))
)

suppressWarnings(
modelsummary(models_hetfx_int, output = "gt",
                coef_map = c('treat1' = 'Treatment', 'imm_1' = 'Immigration',
'treat1:imm_1' = 'Treatment x Immigration',
            '(Intercept)' = 'Intercept'),
            gof_omit = "BIC|AIC|R2 Adj.|F|RMSE|Log.Lik.",
             stars = c('*'=.1, "**"=.05, "***"=.01),
             add_rows = tribble(~term,  ~Base,  ~Base, ~Base,
               "Weighted", "No", "No", "No",
               "HC3 Robust SEs", "Yes", "Yes", "Yes",
               "Survey Robust SEs", "No", "No", "No",
             ))
)
ggplot(spain, aes(x=age, y=weightbins)) +
  geom_jitter(height=.2, width=.2, alpha = .5, na.rm=TRUE) +
  ylab("Weight Bin") +
  theme_bw()
# standard error function:
stderr <- function(x, na.rm=FALSE) {
  if (na.rm) x <- na.omit(x)
  sqrt(var(x)/length(x))
}
```

```r
# aggregate up imm_1 by age, Spain:
age_imm_es <- spain %>%
  filter(!is.na(age)) %>%
  group_by(age) %>%
  summarise(imm_1_mean = mean(imm_1, na.rm=TRUE),
            imm_1_se = stderr(imm_1, na.rm=TRUE),
            count = n()) %>%
  ungroup()


# plot:
ggplot(age_imm_es, aes(x = age, y = imm_1_mean)) +
  geom_point() +
   geom_errorbar(aes(ymin = imm_1_mean - 1.96 * imm_1_se, ymax = imm_1_mean +
1.96 * imm_1_se), width = 0) +
  ylab("Average imm_1 (higher more positive)")+
  xlab("Age") +
  theme_minimal()

# use the age categories already defined in the replication data:
models_hetfxage <- list(
  'Age < 25' = lm(support ~ treat, data=spain[spain$agecat==1,]),
  'Age 25-34' = lm(support ~ treat, data=spain[spain$agecat==2,]),
  'Age 35-44' = lm(support ~ treat, data=spain[spain$agecat==3,]),
  'Age 45-54' = lm(support ~ treat, data=spain[spain$agecat==4,]),
  'Age 55-64' = lm(support ~ treat, data=spain[spain$agecat==5,]),
  'Age >64' = lm(support ~ treat, data=spain[spain$agecat==6,])
)
modelsummary(models_hetfxage, output = "gt", stars = TRUE, robust = TRUE)
# aggregate up imm_1 by age, UK:
age_imm_uk <- uk %>%
  filter(!is.na(age)) %>%
  group_by(age) %>%
  summarise(imm_1_mean = mean(imm_1, na.rm=TRUE),
            imm_1_se = stderr(imm_1, na.rm=TRUE),
            count = n()) %>%
  ungroup()


# plot:
ggplot(age_imm_uk, aes(x = age, y = imm_1_mean)) +
  geom_point() +
   geom_errorbar(aes(ymin = imm_1_mean - 1.96 * imm_1_se, ymax = imm_1_mean +
1.96 * imm_1_se), width = 0) +
  ylab("Average imm_1 (higher more positive)")+
  xlab("Age") +
  theme_minimal()
```

```r
# use the age categories already defined in the replication data:
models_hetfxage_uk <- list(
  'Age < 25' = lm(support ~ treat, data=uk[uk$agecat=="18-24",]),
  'Age 25-34' = lm(support ~ treat, data=uk[uk$agecat=="25-34",]),
  'Age 35-44' = lm(support ~ treat, data=uk[uk$agecat=="35-44",]),
  'Age 45-54' = lm(support ~ treat, data=uk[uk$agecat=="45-54",]),
  'Age 55-64' = lm(support ~ treat, data=uk[uk$agecat=="55-64",]),
  'Age >64' = lm(support ~ treat, data=uk[uk$agecat=="65+",])
)
modelsummary(models_hetfxage_uk, output = "gt", stars = TRUE, robust = TRUE)
# run the linear model without weights
modelES_noweight <- lm(support ~ treat*imm_1, data=spain)
# use lm_robust for robust SEs that can be used by margins(). Set se_type =
"HC3" to be consistent with summ(.,robust=TRUE). Point estimates are of course
numerically identical bar rounding.
modelES_noweight_robust   <-   estimatr::lm_robust(support   ~   treat*imm_1,
data=spain, se_type = "HC3")

# generate code per replication materials, adding in the CI and making SE robust
predES <- interact_plot(modelES_noweight, pred = imm_1, modx = treat, interval
= TRUE, robust = TRUE,
                        colors = colors)+
  labs(title="",
       y="Predicted support for\nLGBT+ education in schools",
       x="")+
  theme_minimal()+
  theme(legend.position = "none",
        axis.text.x =element_blank())+
  annotate(
    geom="text", x = 2.5, y = .65, size = 4, color = "#d11141", fontface=2,
    label = "Slope for\ntreated group")+
  annotate(
    geom = "curve", x =1.6, y = .6, xend = 1.2, yend = .44,
    curvature = .4, arrow = arrow(length = unit(2, "mm")), colour="#d11141")+
  annotate(
    geom="text", x = 6, y = .4, size = 4, color = "#205C8A", fontface=2,
    label = "Slope for\ncontrol group")+
  annotate(
    geom = "curve", x =5, y = .4, xend = 2.6, yend =.31,
    curvature = -.4, arrow = arrow(length = unit(2, "mm")), colour="#205C8A")

gg_df <-
  # move to robust object
  modelES_noweight_robust %>%
  margins(at = list(imm_1 = seq(0, 10, by = 1))) %>%
  summary %>%
  as.data.frame() %>%
  filter(factor == "treat1")
```

```
ameES<- ggplot(gg_df, aes(imm_1, AME)) +
  geom_point(colour="#d11141") +
  geom_line(colour="#d11141") +
  coord_cartesian(xlim = c(0, 10)) +
  # change to 95% ci:
   geom_errorbar(aes(ymax = (AME-SE*1.96), ymin = (AME+SE*1.96)), width = 0,
colour="#d11141") +
  geom_hline(yintercept = 0, linetype = "dashed", colour="#205C8A") +
  xlab("Pre-treatment attitudes towards immigration")+
  ylab("Conditional ATE") +
  theme_minimal()

predES/ameES+
   plot_annotation(title = 'Conditional average treatment effect: Study 2
(Spain)',
               theme = theme(plot.title = element_text(size = 14, face="bold")))
```