A Comment on Turnbull-Dugarte and López Ortega (2024)

Daniel de Kadt
London School of Economics, Department of Methodology
d.n.de-kadt@lse.ac.uk

Introduction

Turnbull-Dugarte and López Ortega (2024) consider whether increasing acceptance of homosexuality in Western countries may be partly attributable to 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 disfavour the ethnic out-group.

The paper presents results from two similar experiments, conducted in the UK ("study 1" in the paper) and Spain ("study 2" in the paper) respectively. In these two experiments respondents were randomly assigned to read vignettes about protests of LGBT+ education in schools. In the UK experiment, the control vignette featured protestors with conventional white-British names, while the treatment vignette featured protestors with conventional Muslim names and a photograph of protestors in identifiably Muslim dress. In the Spain experiment, 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.

The key finding is that being primed with the ethnic out-group (Muslims in both cases) leads to an increase in support for LGBT+ inclusion in schools, but that this effect is generally stronger (in the UK case only present) among those with pre-existing negative attitudes towards immigrants. The authors argue that this heterogeneity is evidence of instrumentalism. Individuals who are pre-disposed to disfavour the out-group are more likely to support LGBT+ inclusion in schools when they see that support as in opposition or contrast to an ethnic out-group.

In this document I outline some concerns with the paper. Two are major concerns:

1. The published results of the Spain experiment rely on the use of post-stratification survey weights in the regression analyses. This choice is inconsistent with the analysis of the UK experiment, where no such weights are used. When data are not weighted, the results in the Spain experiment are largely null, both statistically and substantively. This sensitivity appears to derive from the weights having an unusual distribution, with roughly two-thirds of the data receiving weights less than 0.1, and roughly one-third receiving weights of approximately 3. The choice to use weights in the Spain experiment is never explained, and the manner in which the weights were created is not documented with any detail 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. Further, 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. It appears that a small segment of the data – those aged 45 and older – drive the results in the Spain experiment. No such age heterogeneities appear in the UK experiment.

2. Heteroskedasticity-robust standard errors are used inconsistently in the regression analyses throughout. Robust standard errors are presented in one visualisation (Figure 4, albeit only partially) and two tables (Tables A8 and A11), yet they are not included in others (e.g. Figures 3, 6, 7, 8, and Tables A7, A9, A10). The choice to selectively not present robust standard errors is not documented in the paper or supplementary materials, is not consistent with best practices, and meaningfully changes the statistical significance of key results in Figure 6, Figure 7, Table A9, and Table A10, which present the main results for the Spain experiment. Further, when the weighted analyses are re-estimated with robust standard errors specifically designed for survey weight regressions, the results are no longer statistically significant at conventional levels.

The remaining more minor concerns relate to inconsistencies between the reporting in the paper and the visualisation of results:

- Figures 3 and 6 in the paper, which present the results of the key interaction analysis. These figures present results from a non-linear logistic regression when the text explicitly states a linear regression was used. This issue also affects Figure 8 though I do not engage that Figure further.
- Figures 4 and 7 in the paper, which present conditional experimental results using a dot-plot. The data points in the figures do not represent the underlying data, but are instead jittered predicted values from a regression. These are not data points in any meaningful sense, and are misleading to readers.
- The absence of confidence intervals in some of the visualisations, and that, where confidence intervals are presented, they are 90% confidence intervals. While in principle an acceptable choice, this is never documented in the paper or supplementary materials, and readers would understandably expect a different default.

I enumerate a few other very minor issues in the Conclusion. Finally, I would note that the paper is largely computationally reproducible without issue, and offer only minor (somewhat opinionated) comments on the replication package in Appendix 1.

The Spain Experiment

The UK and Spain experimental designs are very similar, though the survey platforms used are different and the properties of the samples are thus quite different too. In the UK experiment, the authors use data from a "nationally representative" sample of 1151 complete respondents (details on the origin of the sample, e.g. vendor, population targets, and so forth are not provided in the paper). In the Spain experiment, the authors use a "convenience sample" of 1216 complete respondents through Prolific.

Between these two samples the authors deviate in their analytical choices. In the UK sample the data are analysed without survey weights, using ordinary least squares for the results in supplementary Table A7 and A8 and logistic regression for Figure 3. In the Spain sample the data are analysed with weighted least squares for the results in supplementary Table A9 and A10 and weighted logistic regression for Figure 6. The weights used are seemingly post-stratification survey weights designed to "approximate population parameters based on gender, age, education, and geographical region" (Turnbull-Dugarte and López Ortega 2024, 1370).

This choice is not explained in the paper. Conventional best practices in the discipline generally advocate for not weighting convenience samples (Mullinix et al. (2015), Franco et al. (2017), Miratrix et al. (2018)), and so this choice is particularly worth noting. Furthermore, information about the exact process or parameters that generated these weights is not available in the paper, supplementary material, or replication material, and the weights were seemingly not provided by the vendor Prolific (they do not ordinarily provide weights).

Re-Analysis of the Spain Experiment

When the data are not weighted and are analysed with OLS or logistic regression as in the UK experiment, the results are markedly different from those reported in the paper. I first reproduce Table A9 from the supplementary materials which presents the key results for the Spain Experiment as Table 1. I then reproduce Table A9 without weighting as Table 2. None of the treatment effect estimates, nor the interaction term, are statistically significant in the Spain experiment when the data are not weighted. In Models 1, 3, and 4, the point estimate is close to zero, while in Model 2 the point estimate is closer to that reported in the paper, though still half the equivalent coefficient in Table A9, and the interaction is close to zero.

A second point in relation to the estimation of Table A9 concerns standard error estimation. The authors are inconsistent in their choices – supplementary Tables A7 (UK), A9 (Spain), and A10 (Spain) do not present heteroskedasticity-robust standard errors, while Tables A8 (UK) and A11 (Spain) do. Similar inconsistencies occur in the Figures. The visual confidence bands in Figure 4 are based on robust standard errors, but the superimposed text-based indications of statistical significance (e.g. "ATE = .10***") do not. Figures 3, 6, and 7 do not report robust standard errors at all. There is no explanation provided for these inconsistencies, and by and large robust standard errors would be recommended in, at the very minimum, the cases with binary dependent variables (e.g. Tables A7 and A9). Table 3 presents the results of Table A9 re-estimated with robust standard errors. The statistical significance of the key coefficients in all four models, but most notably Model 2 (treatment and interaction) and Model 4 (the anti-immigration sub-sample), materially changes due to this choice.

Additionally, 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. Table 4 reproduces Table A9 with corrected survey-weight standard errors (that are also robust to heteroskedasticity), using the `survey::svyglm()` function in `R`. As with the use of robust standard errors, key statistical results for Models 2 and 4 are again materially changed by this choice.

Table 1: Spain Experiment: Regression Results With Weights and Classical (Non-Robust) Standard Errors (Computational Reproduction of Table A9)

	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.238***	0.180***	0.784***	0.484***
	(0.033)	(0.044)	(0.019)	(0.030)
treat1	0.095***	0.211***	0.111***	0.103*
	(0.025)	(0.062)	(0.027)	(0.044)
imm_1	0.065***	0.075***		
	(0.005)	(0.007)		
treat1 × imm_1		−0.019*		
		(0.009)		
Num.Obs.	1196	1196	700	516
R2	0.154	0.157	0.023	0.011
R2 Adj.	0.152	0.155	0.021	0.009
AIC	22739.9	22737.8	14418.3	8629.7
BIC	22760.3	22763.3	14432.0	8642.5
Log.Lik.	−11365.974	−11363.917	−7206.155	−4311.859
F	108.429	73.844	16.356	5.476
RMSE	0.40	0.40	0.32	0.49

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

Table 2: Spain Experiment: Regression Results Without Weights (Re-Estimation of Table A9)

	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.366***	0.331***	0.871***	0.601***
	(0.034)	(0.047)	(0.017)	(0.031)
treat1	0.026	0.094	0.037	0.034
	(0.023)	(0.065)	(0.024)	(0.043)
imm_1	0.059***	0.065***		
	(0.005)	(0.007)		
treat1 × imm_1		−0.010		
		(0.009)		
Num.Obs.	1196	1196	700	516
R2	0.125	0.125	0.004	0.001
R2 Adj.	0.123	0.123	0.002	−0.001
AIC	1163.2	1164.0	371.3	724.7
BIC	1183.6	1189.4	384.9	737.5
Log.Lik.	−577.604	−576.975	−182.635	−359.360
F	84.868	57.008	2.482	0.612
RMSE	0.39	0.39	0.31	0.49

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

Table 3: Spain Experiment: Regression Results With Weights and HC3 Robust Standard Errors (Re-Estimation of Table A9)

	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.238***	0.180*	0.784***	0.484***
	(0.057)	(0.071)	(0.041)	(0.048)
treat1	0.095*	0.211+	0.111*	0.103
	(0.042)	(0.112)	(0.051)	(0.069)
imm_1	0.065***	0.075***		
	(0.008)	(0.011)		
treat1 × imm_1		−0.019		
		(0.016)		
Num.Obs.	1196	1196	700	516
R2	0.154	0.157	0.023	0.011
R2 Adj.	0.152	0.155	0.021	0.009
AIC	22739.9	22737.8	14418.3	8629.7
BIC	22760.3	22763.3	14432.0	8642.5
RMSE	0.40	0.40	0.32	0.49

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

Table 4: Spain Experiment: Regression Results with Weights and Survey-Robust Standard Errors (Re-Estimation of Table A9)

	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.238***	0.180*	0.784***	0.484***
	(0.057)	(0.070)	(0.040)	(0.048)
treat1	0.095*	0.211+	0.111*	0.103
	(0.042)	(0.110)	(0.050)	(0.068)
imm_1	0.065***	0.075***		
	(0.008)	(0.011)		
treat1 × imm_1		−0.019		
		(0.015)		
Num.Obs.	1196	1196	700	516
R2	0.154	0.157	0.023	0.011
R2 Adj.	0.152	0.155	0.021	0.009
AIC	1371.7	1373.0	576.4	813.5
BIC	38517.7	38592.4	30223.4	12532.9
Log.Lik.	−19244.654	−19278.465	−15101.880	−6257.106
F	42.690	30.231	4.875	2.243
RMSE	0.40	0.40	0.32	0.49

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

Similar broad conclusions can be drawn for the ancillary analysis of Western liberal values when weights are removed. Table A11 in the supplementary materials is reproduced without weights as Table 5. 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.

Table 5: Spain Experiment: 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
(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)
treat1	0.508	0.493	−0.242	0.329	−0.245	0.432
	(0.471)	(0.426)	(0.455)	(0.472)	(0.594)	(0.595)
imm_1	0.256***	0.005	0.143**	0.210***	−0.343***	−0.083
	(0.048)	(0.045)	(0.049)	(0.050)	(0.060)	(0.062)
treat1 × imm_1	−0.088	−0.051	0.007	−0.061	0.009	−0.058
	(0.068)	(0.062)	(0.066)	(0.069)	(0.083)	(0.084)
Num.Obs.	1163	1171	1180	1179	1144	1113
R2	0.044	0.003	0.022	0.029	0.068	0.009
R2 Adj.	0.042	0.000	0.019	0.026	0.066	0.006
AIC	5449.1	5258.8	5556.2	5651.6	5873.5	5688.1
BIC	5474.4	5284.1	5581.6	5677.0	5898.7	5713.2
Log.Lik.	−2719.563	−2624.398	−2773.100	−2820.809	−2931.737	−2839.065
F	17.876	0.980		11.497		
RMSE	2.51	2.28	2.54	2.65	3.14	3.10

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

Heterogeneous Effects by Weights

This sensitivity to the use of weights is likely attributable to their unusual distribution, shown in Figure 1. The weights are clustered around two points, with nearly two-thirds (63%) of the data receiving very low weights close to zero (under 0.01), and nearly another third (31%) all receiving a weight of 3.0000919 (rounded to the seventh decimal). Only 6% of observations receive weights greater than 0.01 and less than 3. 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).

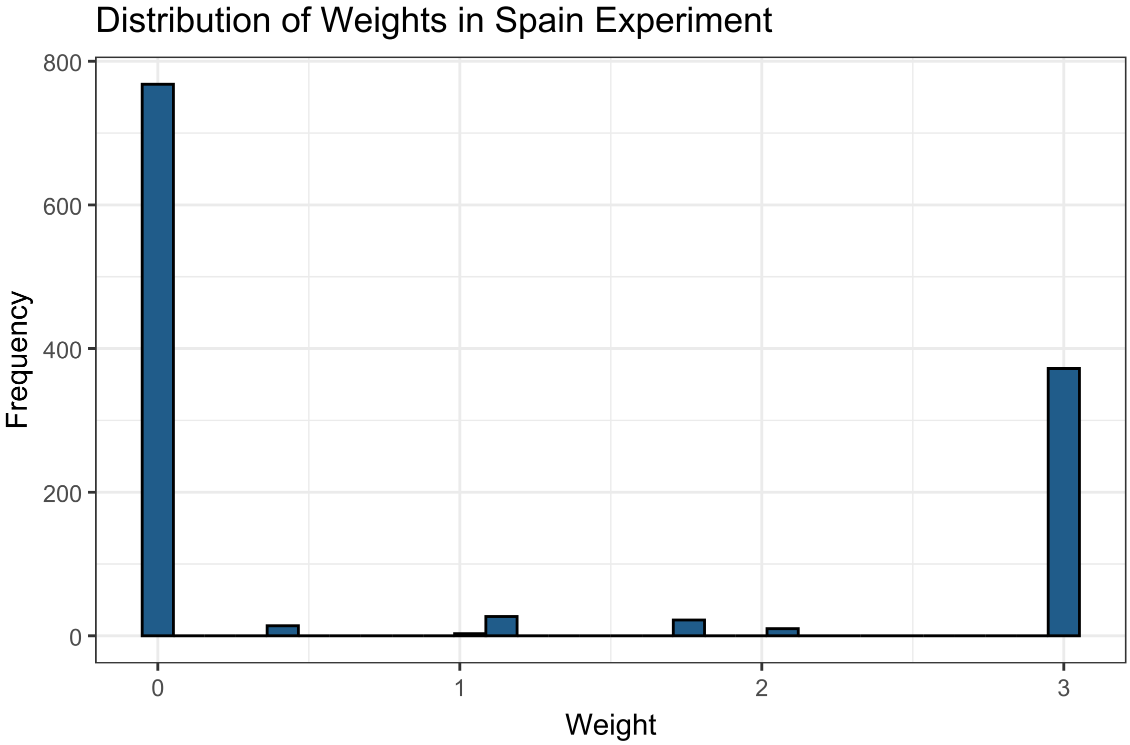


Figure 1: Histogram of Weights in Spain Experiment

I now 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.

Focusing first on only the anti-immigration sub-sample, those theorized to be most likely to respond to the treatment, Table 6 shows that even among these respondents 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.

Table 6: Spain Experiment: Heterogeneous Effects By Weight Bin for Anti-Immigrant Sub-Sample

	Low Weights	Mid Weights	High Weights
(Intercept)	0.687***	0.682***	0.463***
	(0.041)	(0.104)	(0.052)
treat1	−0.018	−0.182	0.132+
	(0.055)	(0.169)	(0.074)
Num.Obs.	294	38	184
R2	0.000	0.034	0.018
R2 Adj.	−0.003	0.007	0.012
AIC	393.4	58.1	269.3
BIC	404.4	63.0	278.9
Log.Lik.	−193.676	−26.067	−131.645
F	0.111	1.257	3.251
RMSE	0.47	0.48	0.49

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

Finally, a similar analysis using all the data, presented in Table 7, 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. These two sets of results are evidence against the authors' hypothesis that the treatment effect is conditional on immigration attitudes – it appears any conditionality in the Spain experiment is in fact with respect to the weights, less immigration attitudes.

Table 7: Spain Experiment: Heterogeneous Interaction Effects By Weight Bin

	Low Weights	Mid Weights	High Weights
(Intercept)	0.477***	0.487*	0.161*
	(0.077)	(0.206)	(0.074)
treat1	−0.019	−0.159	0.233*
	(0.103)	(0.336)	(0.118)
imm_1	0.050***	0.049+	0.075***
	(0.010)	(0.025)	(0.011)
treat1 × imm_1	0.001	0.017	−0.019
	(0.013)	(0.045)	(0.017)
Num.Obs.	755	76	365
R2	0.096	0.122	0.164
R2 Adj.	0.092	0.085	0.157
AIC	640.9	82.8	424.2
BIC	664.1	94.4	443.7
Log.Lik.	−315.474	−36.384	−207.105
F	26.569	3.323	23.568
RMSE	0.37	0.39	0.43

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

Who is Being Up-Weighted?

The authors devised the weights to match Spanish population parameters in terms of gender, age, education, and region. We can assess which characteristics are being up-weighted by examining the characteristics of a range of covariates across the weight bins, shown in Table 8. 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 are strongly up-weighted, while those who are younger (the bulk of the data) are heavily down-weighted. As a result, 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.

Table 8: Spain Experiment: 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

A more detailed visualisation of the distribution of age across weight bins is presented in Figure 2. 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.

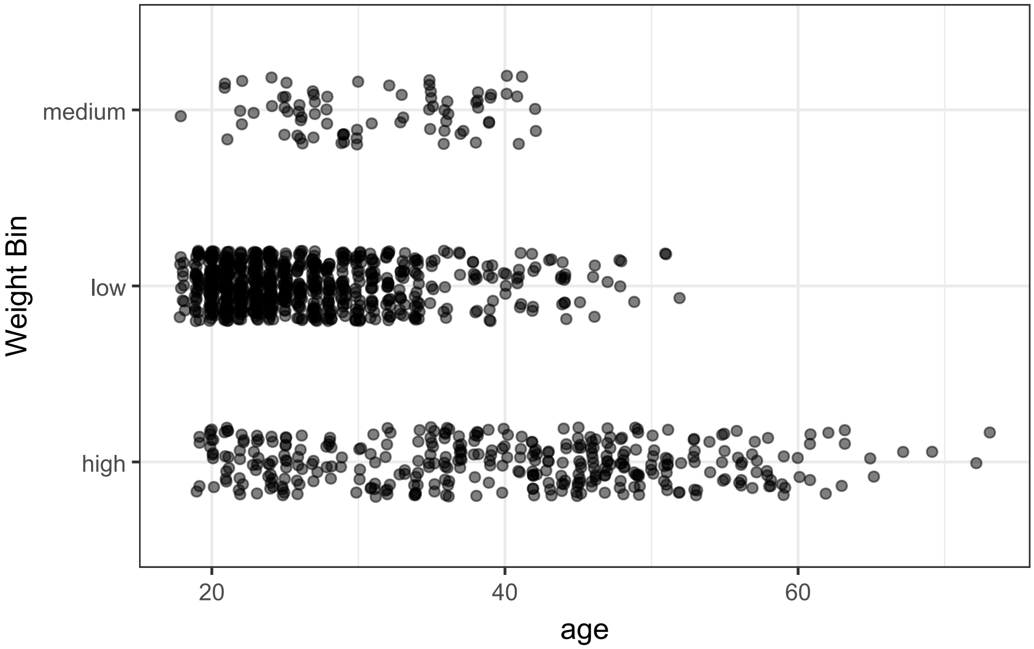


Figure 2: Spain Experiment: Age Distributions by Weight Bin

Table 9 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 roughly 150 observations in the entire sample 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 9: Spain Experiment: 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**
	(0.024)	(0.027)	(0.047)	(0.068)	(0.104)	(0.186)
treat1	−0.039	0.004	−0.035	0.367***	0.131	0.194
	(0.034)	(0.039)	(0.071)	(0.090)	(0.147)	(0.231)
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

Do Similar Age Effects Occur in the UK Experiment?

It appears that the results in the Spain experiment 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.

Weighing against this deeper interpretation of these results is the fact that in the UK Experiment there are no coherent age heterogeneities. The results of the same analysis in the UK experiment are shown in Table 10. 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 10: UK Experiment: 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***
	(0.068)	(0.051)	(0.043)	(0.046)	(0.048)	(0.045)
treat1	0.189*	−0.149*	0.094	−0.038	0.073	−0.013
	(0.086)	(0.073)	(0.063)	(0.065)	(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

Visusalisations

I now turn to some more minor concerns. I focus primarily on the key visualisations in the paper, but I make a number of additional comments throughout.

The key evidence in the paper comes from a series of interaction specifications studying how the experimental interventions affect stated support for LGBT+ education in schools, conditional on pre-treatment immigration attitudes. The authors anticipate that those respondents who are already negatively disposed toward immigration will show larger effects than those who are not. These interacted results are presented through two primary visualisation techniques in the paper, for both experiments, while all 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 attitudes towards immigration. The second set, Figures 4 and 7 respectively, are dot-plots showing the means and confidence intervals for each condition in two subgroups – those who are pro-immigration (defined as higher than the mean), and those who are anti-immigration (lower than the mean). I address both sets of figures below, in turn.

Figures 3 and 6

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` might be referred to as the treatment variable and `imm_1`, an 11-point scale variable, as the moderator. Note that there is likely a typographical error in the formalization 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. 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. Instead the plots are generated with the `jtools::interact_plot()` function in `R` on the basis of an underlying regression implemented in `R` 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. In the notes of Figures 3 and 6 the authors direct readers to supplementary material tables A7 and A9 respectively for "full regression output" – these tables actually 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 a function of the use of logistic regression. The 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 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 (conducted by author, but not reported in this document).

Finally, a note on the presentation of uncertainty in these figures. First, the authors do not include confidence bands for the top panel of the figure – the reason for this omission is unclear. However, in the lower panel the authors present only 90% confidence intervals. This detail is not mentioned anywhere in the paper or the figures, and was only discernible in the code by examining the hard-coded critical values used to generate the confidence intervals. As noted previously, in neither plot are these confidence intervals are not based on robust standard errors. 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.

For completeness, I reproduce Figures 3 and 6 below, with two corrections, as Figure 3 and Figure 4. 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 the logistic regression approach, 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. Additionally I reproduce Figure 6 as Figure 5, which includes the above corrections but removes the weights. I do not reproduce Figure 8, but this figure is also based on logistic regression and the same issues apply.

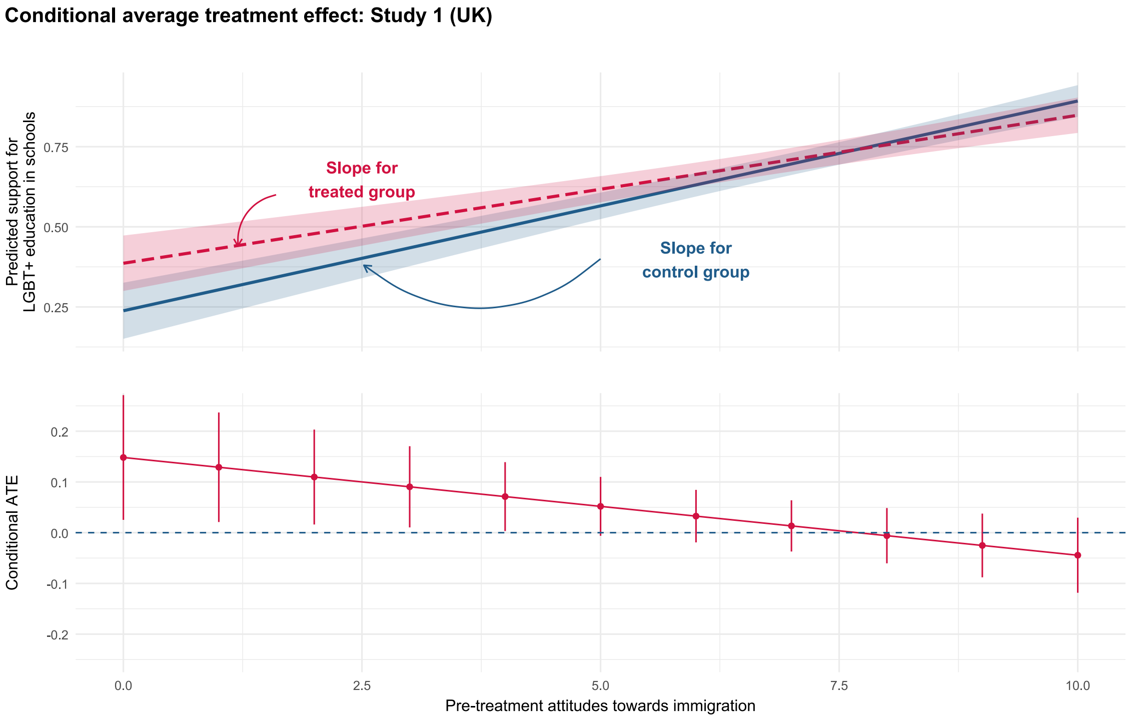

Conditional average treatment effect: Study 1 (UK)

Figure 3: UK Experiment: Corrected Reproduction of Figure 3

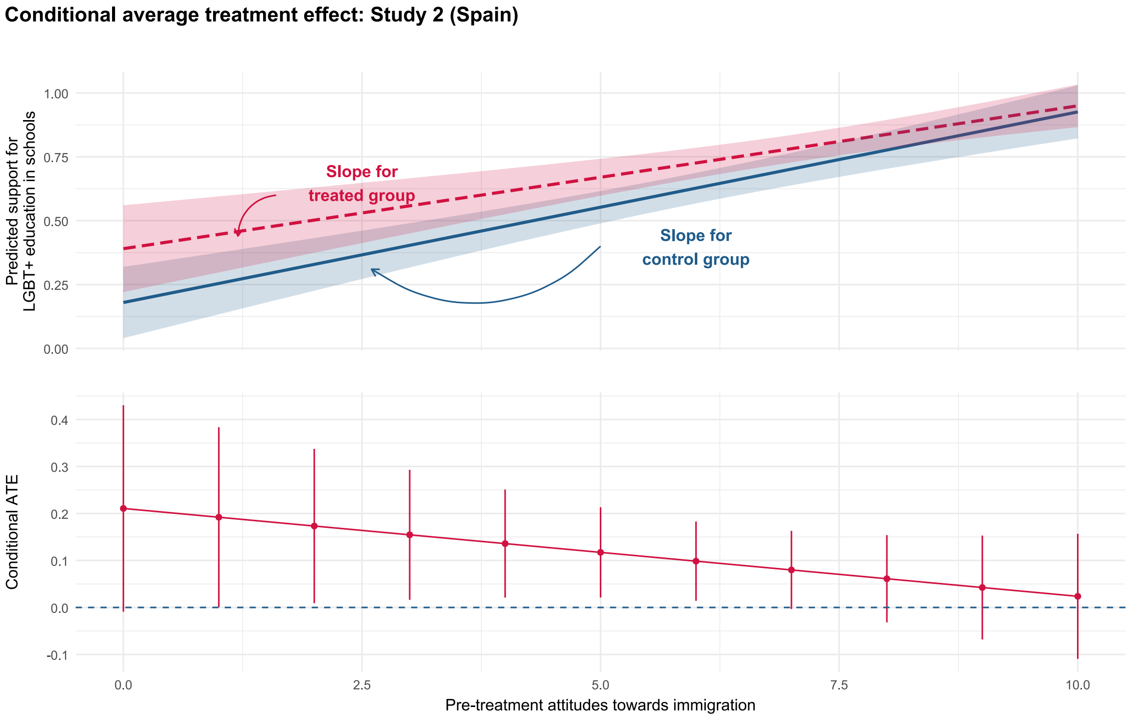

Conditional average treatment effect: Study 2 (Spain)

Figure 4: Spain Experiment: Corrected Reproduction of Figure 6

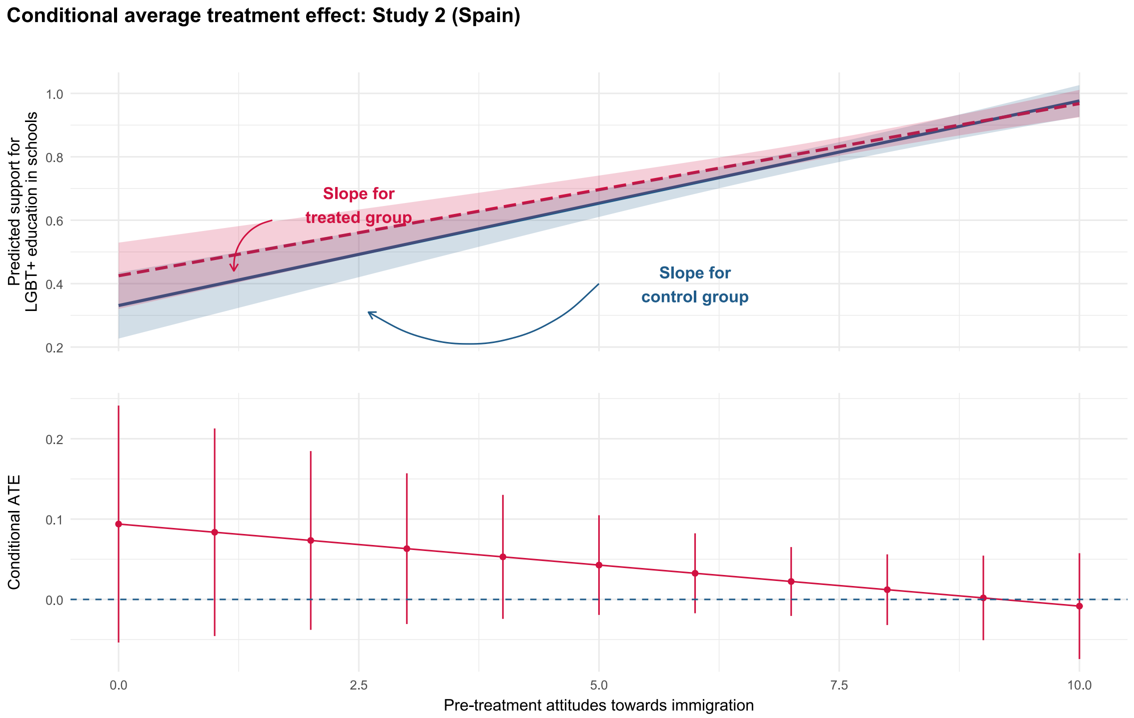


Figure 5: Spain Experiment: Corrected Reproduction of Figure 6 Without Weighting

Figures 4 and 7

The second set of visualisations, Figures 4 and 7 respectively, are jittered (noise-applied) dotplots 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 plots 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.

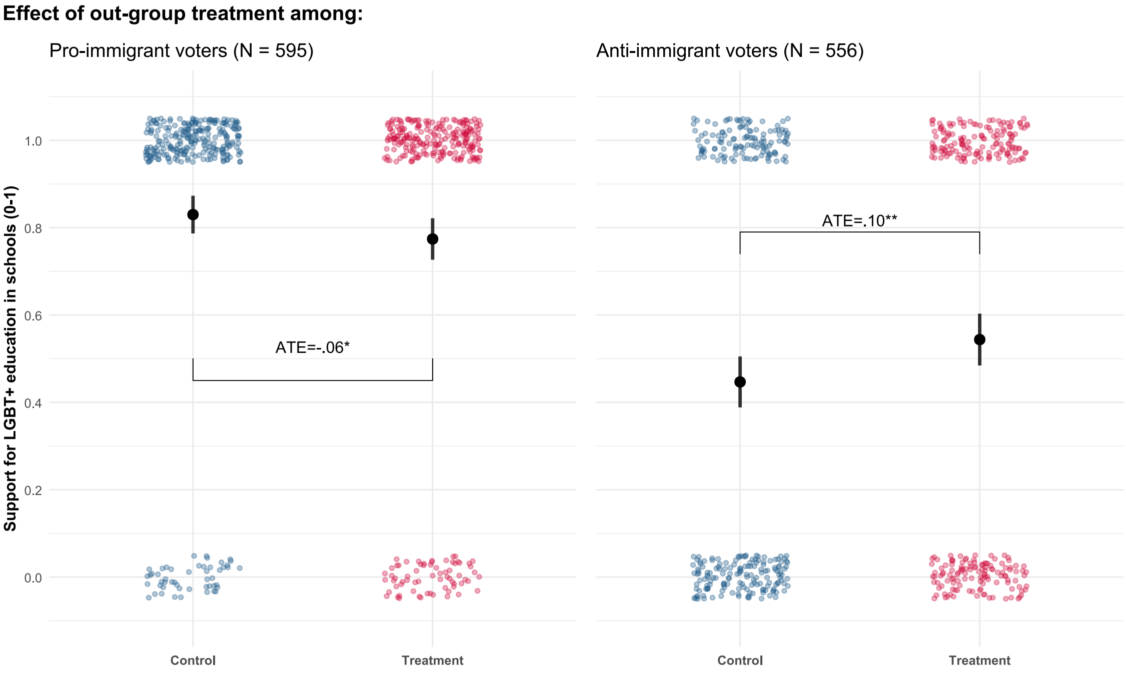
The data points in Figures 4 and 7 do not represent the underlying data on which the analyses are based. Each data point's value on the y-axis is not that data point's value on the binary outcome variable Y or the continuous variable on which the dichotomization was based. Instead they are the fitted (or "predicted") value $\hat{Y}$ of the binary outcome for each observation, given the underlying analysis 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. Further, the original 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.

Below I reproduce Figures 4 and 7 with a number of corrections. The figures below show the (jittered) data points as they feature in the regression analysis, and the confidence intervals are the 95% confidence intervals based on robust standard errors. The text representations of point estimates and statistical significance are updated to reflect estimation with robust standard errors. In the UK experiment the figures do not much alter the substantive conclusions of the authors, though the confidence intervals widen and the p-values do increase across the board. In the Spain experiment the changes are more notable: with robust standard errors the anti-immigrant plot shows no statistically significant effect (even while weighting).

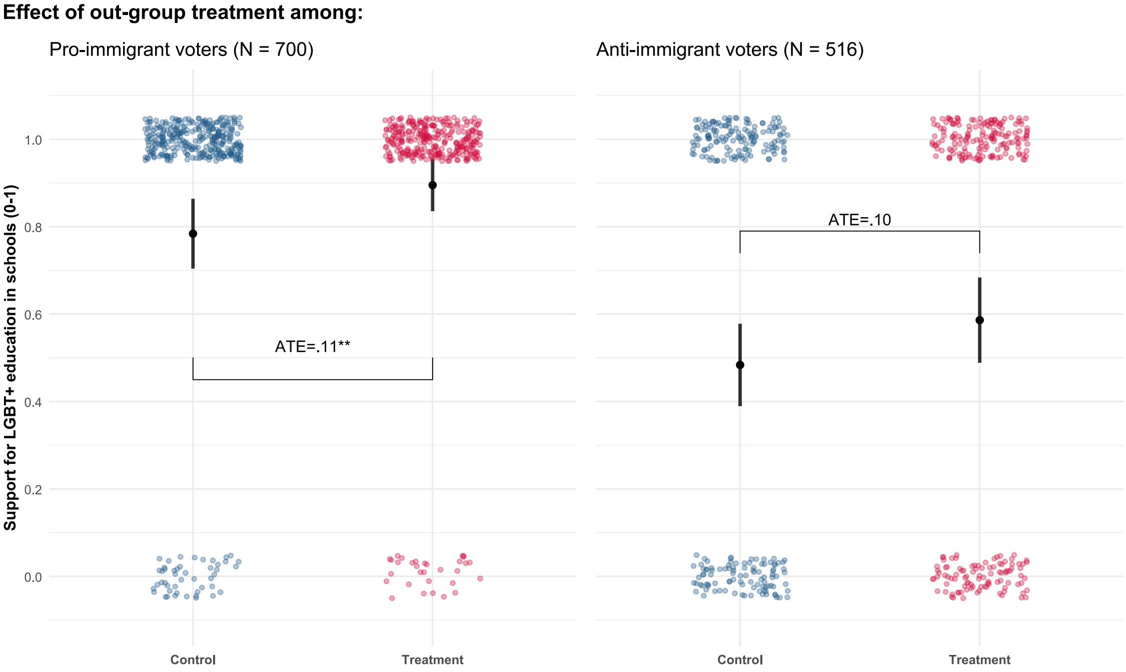
Further, the updated plots do reveal a broader point of concern with specifically the pro-immigrant samples, most acutely in the Spain sample but also in the UK sample: there are very few respondents who take a zero value on the dependent variable in either the treatment or control conditions. In the UK sample across both conditions the proportion of the pro-immigrant subsample that takes a zero value on the dependent variable is 0.2. In the Spain sample that proportion is just 0.11.

In both cases this raises concerns about potential ceiling effects that might drive 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 simply a ceiling effect.



Lines are 95% confidence intervals.Treatment group outcome statistically distinct at p<.1(*), p<0.05(**), & p<0.01(***)

Figure 6: UK Experiment: Corrected Reproduction of Figure 4



Lines are 95% confidence intervals. Treatment group outcome statistically distinct at p<.1(*), p<0.05(**), & p<0.01(***)

Figure 7: Spain Experiment: Corrected Reproduction of Figure 7

Conclusion

In this document I have highlighted two main concerns with Turnbull-Dugarte and López Ortega (2024):

1. 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 the Spain experiment are largely null, both statistically and substantively. This sensitivity appears to be driven by the weights' 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. It appears that a small segment of the data – those aged 45 and older – drive the results in the Spain experiment. Similar age-based heterogeneities do not appear in the UK experiment.

2. 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 Spain experiment (regardless of weighting), presented in Figure 6, Figure 7, Table A9, and Table A10.

I have also noted a number of additional but more minor issues, most pressingly:

- Figures 3 and 6 in the published paper present results from a non-linear logistic regression when the text explicitly states a linear regression was used. The results in supplementary tables A7 and A9, which are implied to be the same as those in the figures, are from a linear regression. Fortunately, after correcting the figures the substantive conclusions drawn do not change much, but the discrepancy is concerning.

- Figures 4 and 7 in the published paper present jittered predicted values from a regression. These are not data points in any meaningful sense, and are misleading to readers. Corrected figures reveal the possibility of ceiling effects for the pro-immigration sample, which could explain the heterogeneous effects that are interpreted as dispositive evidence in favour of the authors' theory.

- The lack of confidence intervals in some visualisations, and the use of 90% confidence intervals in others, a choice that is not documented in the paper or supplementary materials.

There are also a few much more minor points, not all of which I have not explicitly addressed above:

- The specification of Model 1 in Tables A7 and A9 and in Tables A8 and A10 is inconsistent. In the UK experiment (A7/A8) this "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. This inconsistency is not documented or explained in the paper or supplementary materials, though it does not appear to have much bearing on the results.

- The dichotomizing of the dependent variables is inconsistent. The authors differ in their approach to dichotomization between the main dependent variable (dichotomized at the mean on an 11 point scale) and the ancillary dependent variable (dichotomized at 5 on an 11 point scale). This inconsistency is not documented or explained in the paper or supplementary materials, though it does not appear to have much bearing on the results.
- A typographical error in the regression equation in the published paper.
- The apparent misspecification, or at a minimum lack of sufficient explanation and interpretation, of the linearity test presented in the appendix.

Unfortunately, because the study was seemingly not pre-registered (this is not mentioned in the paper or the supplementary materials) it is very hard to make sense of the heterogeneities that seem to be driving the results.

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.

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: Computational Reproducibility

The paper is almost perfectly computationally reproducible. 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. I note a few minor issues encountered below:

- 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 replicate. 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: All Code Used In This Report

```r
# 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)

# 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
```

```r
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)
# build tables -- linear regression with no weights:
models_rep <- list(
  'Model 1' = lm(support ~ treat + imm_1, data=spain, weight = nationalweight),
  'Model 2' = lm(support ~ treat*imm_1, data=spain, weight = nationalweight),
  'Model 3' = lm(support ~ treat, data=proimmES, weight = nationalweight),
  'Model 4' = lm(support ~ treat, data=noproimmES, weight = nationalweight)
)

# build tables -- linear regression with no weights:
models_noweight <- list(
  'Model 1' = lm(support ~ treat + imm_1, data=spain),
  'Model 2' = lm(support ~ treat*imm_1, data=spain),
  'Model 3' = lm(support ~ treat, data=proimmES),
  'Model 4' = lm(support ~ treat, data=noproimmES)
)

# reproduce Table A9 but with weights, with lm, and with robust SEs (HC3 to be
consistent with summ()):
models_weigh_lm_robust <- list(
    'Model 1' = lm_robust(support ~ treat + imm_1, data=spain, weight =
nationalweight, se_type = "HC3"),
    'Model 2' = lm_robust(support ~ treat*imm_1, data=spain, weight =
nationalweight, se_type = "HC3"),
  'Model 3' = lm_robust(support ~ treat, data=proimmES, weight = nationalweight,
se_type = "HC3"),
  'Model 4' = lm_robust(support ~ treat, data=noproimmES, weight = nationalweight,
se_type = "HC3")
)

# reproduce Table A9 but use svyglm instead:
models_svm <- list(
    'Model 1' = svyglm(support ~ treat + imm_1, design=svydesign(ids=~1,
weights=~nationalweight, data=spain)),
     'Model 2' = svyglm(support ~ treat*imm_1, design=svydesign(ids=~1,
weights=~nationalweight, data=spain)),
      'Model 3' = svyglm(support ~ treat, design=svydesign(ids=~1,
weights=~nationalweight, data=proimmES)),
      'Model 4' = svyglm(support ~ treat, design=svydesign(ids=~1,
weights=~nationalweight, data=noproimmES))
)
modelsummary(models_rep, output = "gt", stars = TRUE)
modelsummary(models_noweight, output = "gt", stars = TRUE)
modelsummary(models_weigh_lm_robust, output = "gt", stars = TRUE)
suppressWarnings(modelsummary(models_svm, output = "gt", stars = TRUE, robust =
TRUE))
mech <- list(
  'EU norms' = lm(pride_valoresUE ~ treat*imm_1, data=spain),
  'Western liberal values' = lm(pride_libertadOCC ~ treat*imm_1, data=spain),
  'Green politics' = lm(pride_verde ~ treat*imm_1, data=spain),
  'Domestic violence protections' = lm(pride_viomach ~ treat*imm_1, data=spain),
  'Spanish flag' = lm(pride_bandera ~ treat*imm_1, data=spain),
  'Spanish military efforts' = lm(pride_mili ~ treat*imm_1, data=spain)
)
modelsummary(mech, output = "gt", stars = TRUE, robust = TRUE)
# create weights plot
weights_plot <- ggplot(spain, aes(x=nationalweight)) +
  geom_histogram(bins=30, fill="#205C8A", color="black") +
  theme_bw() +
  labs(title = "Distribution of Weights in Spain Experiment",
       x = "Weight",
       y = "Frequency")

weights_plot
# 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 - hetfx
models_hetfx <- list(
              'Low       Weights'       =       lm(support       ~       treat,
data=noproimmES[noproimmES$weightbins=="low",]),
              'Mid       Weights'       =       lm(support       ~       treat,
data=noproimmES[noproimmES$weightbins=="medium",]),
              'High       Weights'       =       lm(support       ~       treat,
data=noproimmES[noproimmES$weightbins=="high",])
)

# build tables - interaction hetfx
models_hetfx_int <- list(
            'Low      Weights'      =      lm(support      ~      treat*imm_1,
data=spain[spain$weightbins=="low",]),
            'Mid      Weights'      =      lm(support      ~      treat*imm_1,
data=spain[spain$weightbins=="medium",]),
            'High      Weights'      =      lm(support      ~      treat*imm_1,
data=spain[spain$weightbins=="high",])
)
modelsummary(models_hetfx, output = "gt", stars = TRUE, robust = TRUE)
modelsummary(models_hetfx_int, output = "gt", stars = TRUE, robust = TRUE)
# create a balance table where we show the mean of imm_1, age, edu, gender, and
other covariates by weightbin:
bal_table <- spain %>%
  mutate(gender = as.numeric(as.character(gender)) - 1) %>%
  group_by(weightbins) %>%
  summarise_at(
    c("age", "gender", "edu", "child", "foreignborn", "queer", "imm_1"),
    ~round(mean(as.numeric(as.character(.)), na.rm = T),2)
  ) %>%
# transpose the table so the weightbins are the columns, and the covariates the
rows:
  pivot_longer(cols = c(age, gender, edu, child, foreignborn, queer, imm_1)) %>%
  pivot_wider(names_from = weightbins, values_from = value) %>%
  relocate(low, .before = high) %>%
  rename(`Low Weight Bin` = low, `Medium Weight Bin` = medium, `High Weight Bin`
= high, `Variable` = name)
gt(bal_table)
ggplot(spain, aes(x=age, y=weightbins)) +
  geom_jitter(height=.2, width=.2, alpha = .5, na.rm=TRUE) +
  ylab("Weight Bin") +
  theme_bw()
# use the age categories already defined in the replication data:
models_hetfxage <- list(
  'Age < 25' = lm(support ~ treat, data=spain[spain$agecat==1,]),
```

```r
  '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$agecat==6,])
)
modelsummary(models_hetfxage, output = "gt", stars = TRUE, robust = TRUE)
# use the age categories already defined in the replication data:
models_hetfxage_uk <- list(
  'Age < 25' = lm(support ~ treat, data=uk[uk$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)
# 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")+
```

```r
  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")))
# 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")))
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")))
```