CHAPTER 2

PRECISION IN SURVEY EXPERIMENTS – A NEW METHOD TO IMPROVE BLOCKING ON ORDINAL VARIABLES

2.1 Introduction

Survey experiments collect background information and attempt to uncover treatment effects on public opinion and/or behavior. In order to identify such potential effects, the treatment groups need to be comparable. All treatment groups need to look the same in every measure, i.e. they must be balanced. This can be achieved through random assignment of participants to treatment groups. Randomization, i.e. flipping a coin to decide which treatment group a participant is assigned to, probabilistically results in balance based on the Law of Large Numbers (Urdan, 2010). For small samples, however, it can lead to serious imbalance. It can easily be that the treatment groups will not look the

same. This can leave experimental results in statistically murky waters (Fox, 2015; Imai, 2018; King, Keohane, & Verba, 1994). In survey experiments, the overall sample size is often split across several treatment groups, which can exacerbate the problem. Chong & Druckman (2007), for instance, split 869 participants in a framing experiment on urban growth over 17 treatment groups, which leads to an average of just over 50 participants per group. Randomization is unlikely to lead to balanced treatment groups of this size. Researchers need to employ statistical methods to obtain balanced groups here. Blocking, i.e. arranging participants in groups that are equal in terms of participants' covariates and using random allocation within these groups, can alleviate such worries.

Blocking depends on covariates. In political science, many covariates with high predictive power are categorical variables, i.e. variables where the data can be divided into groups. These include interval (ordered and evenly spaced, e.g. `Income`) and ordinal (ordered and unevenly spaced, e.g. `Education`) variables. To block, these variables are often made numeric, e.g. by assigning the numbers 1-4 to the variable categories. This is acceptable for interval variables as the evenly spaced numbers correspond to the evenly spaced categories. For ordinal variables, however, this can be problematic. An arbitrary evenly spaced string of numbers does not correspond to the unevenly spaced ordinal categories and may misrepresent the data. I propose an ordered probit threshold approach to circumvent this problem: This approach estimates an assumed underlying latent continuous structure underneath ordinal variables whose data-driven categories can then be used for blocking. By training a linear model on meaningful data, it creates numerical thresholds which partition the variable into regions corresponding to the ordinal categories and bins the observations between these thresholds according to the explanatory variables. These binned cases determine which of the original categories make sense given the underlying latent continuous structure. The result is a data-based and non-arbitrary re-estimated set of variable categories. Because of their data-driven estimation, these categories can be

safely used for blocking. This approach allows researchers to block on ordinal variables in survey experiments without making unwarranted assumptions in terms of arbitrary numeric values whilst fully utilizing the ordinal information provided and respecting uneven spaces.

The following sections provide a background on survey experiments and blocking, describe the key aspects of ordinal variables, and outline my proposed ordered probit approach. I then demonstrate the benefits and implications of this approach with external survey data and original data from an online survey experiment. Since there currently is no available tool to block in online survey experiments, I create my own survey environment in `shiny`, which will be described in more detail below.

2.2 Theory

2.2.1 Preliminary Notations on Survey Experiments

The simplest of survey experiments has two potential outcomes for participants i, y_{1i} and y_{0i}, with 1 denoting the treatment and 0 referring to the control. Consider a simplified version of a famous survey experiment by Tversky & Kahneman (1981), where researchers want to test the effect of the mortality format on participants' choices. They provide participants with the following scenario:

> Imagine that the US is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. A program to combat the disease has been proposed. Assume that the exact scientific estimates of the consequences of the program are as follows...

Participants in the control group receive the program description in survival format:

> If the program is adopted, 200 out of 600 people will live.

Participants in the treatment group receive the program description in mortality format:

> If the program is adopted, 400 out of 600 people will die.

All participants are subsequently asked whether they support or oppose the program. The treatment effect for each individual participant i is given by $y_{1i} - y_{0i}$. If both groups of participants look the same regarding their covariates (`Age`, `Education`, `Income` etc.), a comparison of the groups' average support reveals the Average Treatment Effect (ATE) across all participants, $\mathbf{E}[\delta] = \mathbf{E}[y_{1i} - y_{0i}]$. A central characteristic of such a comparison is the fundamental problem of causal inference (Holland, 1986; Rubin, 1974): We are unable to observe both potential outcomes for the same participant at once. In our case, we cannot observe how much participant A supports the program if given the survival format whilst also observing how much the same participant A would have supported the program if given the mortality format. If we could, it would be simple to calculate the true average treatment effect, $\mathbf{E}[\delta] = \mathbf{E}[y_{1i}|T = 1] - \mathbf{E}[y_{0i}|T = 0]$, with $T = 0$ denoting the control and $T = 1$ the treatment group. Since the true average treatment effect is unobservable, we need to use statistical means to assess the counterfactuals. This can be done by balancing the treatment and control groups. If both groups of participants look the same in every measure, we can use the participants who received the mortality format (treatment) to estimate what would have happened to the participants who did not receive the mortality format (control). The crucial aspect is whether the two groups do indeed look the same in terms of participants' covariates. The potential outcome of the control needs to mirror what would have happened in the case of treatment, and vice versa. There are two main means by which this may be achieved: Randomization and blocking.

2.2.2 Randomization

Randomization is equivalent to flipping a coin for each participant to be assigned to treatment or control. This chance procedure gives each participant an equal chance of being assigned to either group (or groups, in case of multiple treatment groups) (Lachin, 1988). Randomization increases covariate balance as the number of participants, n, in-

creases (Imai, King, & Nall, 2009). The larger a researcher's sample, the better the resulting balance from randomization in expectation. Probabilistically, randomization enables the comparison of the average treatment effect to be unbiased, which allows the researcher to attribute any treatment effects to the treatment (King et al., 2007).

While randomization thus guarantees balance as the sample size reaches infinity, it often does not do so in the naturally finite sample sizes researchers actually work with. With huge samples, the Law of Large Numbers predicts that treatment groups selected through randomization will be balanced. With small samples, however, it is possible to get unlucky and end up with unbalanced groups (Imai, King, & Elizabeth A. Stuart, 2008). Blocking can help achieve balance in such scenarios (Epstein & King, 2002).

2.2.3 Blocking

Identical levels in terms of covariates across treatment groups represent the key aspect in experimental studies. In randomization, this is achieved by random chance. In blocking, this is achieved by combining covariate information about the participants with randomization. Specifically, participants are blocked into treatment groups that are similar to one another in terms of the their covariates before treatment is assigned. Their similarity is estimated with the Mahalanobis or Euclidian distance. Blocking is better suited to achieving balance in finite samples than randomization, as it "directly controls the estimation error due to differing levels of observed covariates in the treatment and control groups" (Moore, 2012, p. 463). This is particularly relevant with small samples and a high number of treatment groups, as the overall number of participants needs to be divided up. Figures 2.1 and 2.2 show this visually. A numeric discrete variable with levels 1 to 5 is randomized and blocked for different sample sizes and numbers of treatment groups. This is repeated 100 times for each sample size. Figure 2.1 shows the maximum distances between treatment groups across these repetitions for sample sizes up to 1,000 for two, three, five, and ten treatment groups. Blocking outperforms randomization in

every scenario. The difference between the two methods is smallest for large samples and a small number of treatment groups. For $n = 998$ and two treatment groups, the largest

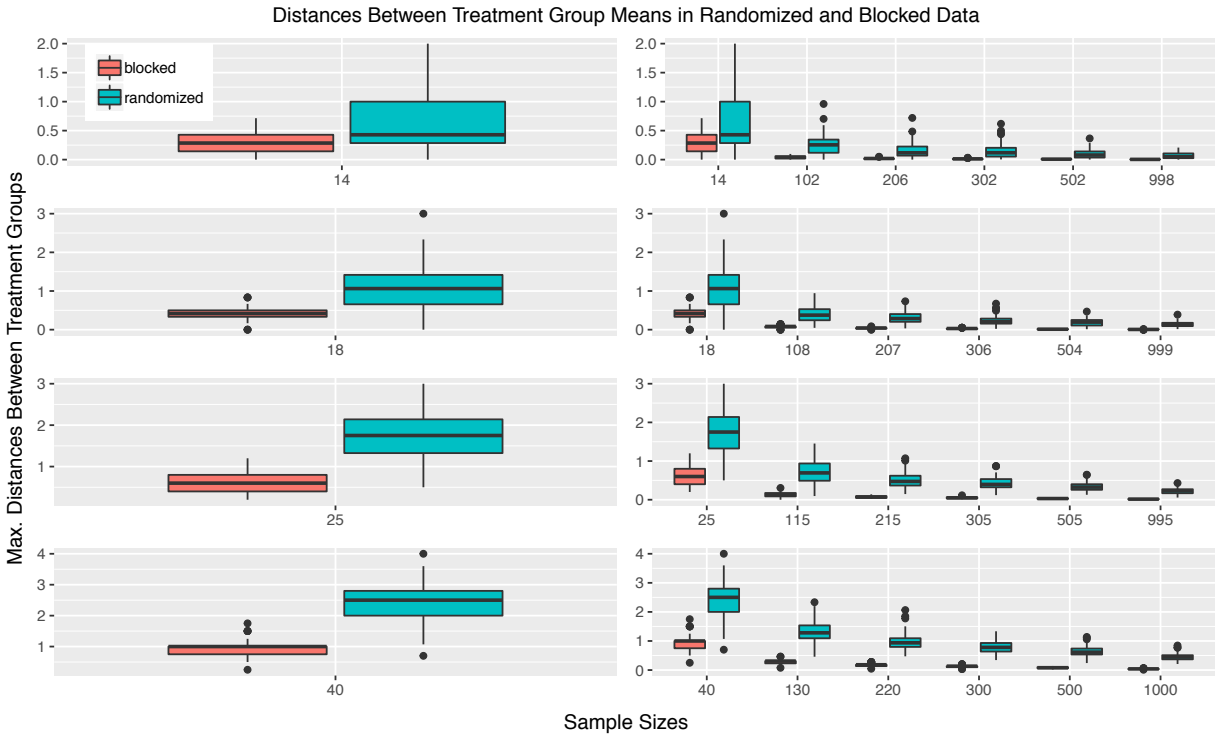

Figure 2.1. Imbalances between treatment groups for blocked and randomized data with increasing sample size for 2 (top row), 3 (second row), 5 (third row), and 10 treatment groups (bottom row). Leftmost pair on each right panel is exactly the pair in the left panel

distance between randomized treatment groups is 0.208, and the largest distance between blocked treatment groups is 0.01. For small samples and a large number of treatment groups, however, the difference is much starker. For $n = 40$ and ten treatment groups, the largest distance between randomized treatment groups is 4, and the largest distance between blocked treatment groups is 1.75. Figure 2.2 shows the distribution of these imbalances.

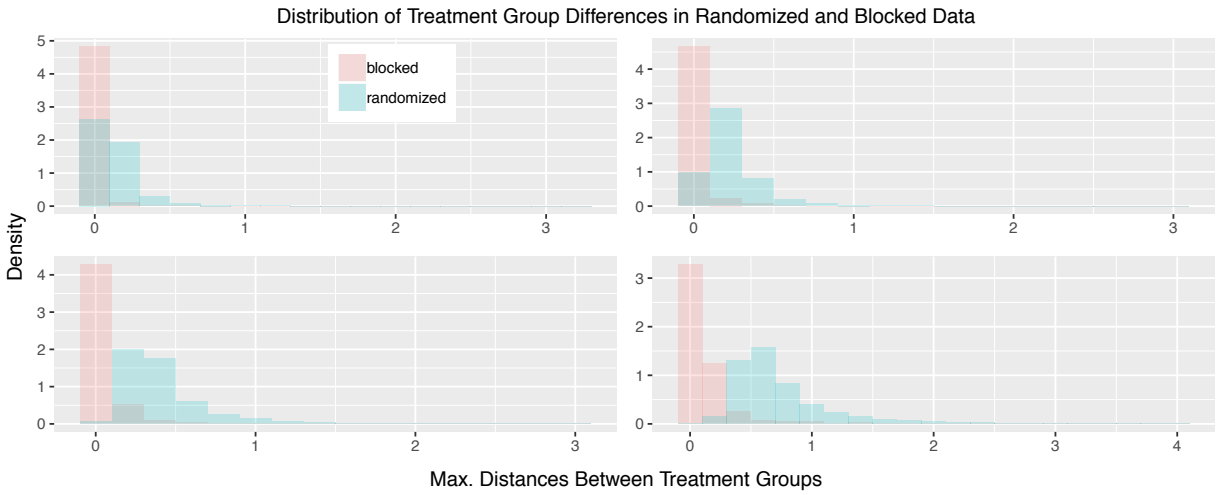

Figure 2.2. Distribution of imbalances between treatment groups for the same blocked and randomized data for 2 (top left), 3 (top right), 5 (bottom left), and 10 (bottom right) treatment groups

Blocking On The Go

In political science, researchers often have an already-collected data set in front of them. One example would be the American National Election Studies (ANES), a pre-existing survey database, which is often used to analyze voter turnout (see for instance Jackman & Spahn, 2018; Leighley & Nagler, 2014), among many others. This setup means all covariate information on all participants is known at the time of assignment, which makes blocking straight-forward. Oftentimes, however, the covariate information of all participants is not known at the time of assignment. This is the case, for instance, for online survey experiments, where each participant completes the survey at differing times. Participants 'trickle in' for treatment assignment as the experiment progresses. 'Traditional' blocking can not be used here, since it relies on covariate information about the entire sample. Instead, we need to block continuously as the experiment progresses, or block 'on the go'. This is called sequential blocking.

Sequential blocking in political science is based on covariate-adaptive randomization, which varies probabilities based on knowledge about previous participants and the current participant (Chow & Chang, 2007). Traditional covariate-adaptive approaches, such as the biased coin design (Efron, 1971) and minimization (Pocock & Simon, 1975), assign the incoming participant to the treatment group with the fewest participants with identical covariate information. This works for discrete covariates as the number of possible covariate levels is finite. For continuous covariates, the number of possible covariate levels rises exponentially. Participants are unlikely to look the same, and identical participants are rare. Blocking on continuous covariates is not possible with these traditional approaches (Eisele, 1995; Markaryan & Rosenberger, 2010; Rosenberger & Lachin, 2002). Moore & Moore (2013) develop a method to do so by exploiting relationships between the current participant's covariate profile and those of all previously assigned participants. They define the similarity between participants with the Mahalanobis distance (MD) between participants q and r with covariate vectors $\boldsymbol{x}_q$ and $\boldsymbol{x}_r$:
$MD_{qr} = \sqrt{(\boldsymbol{x}_q - \boldsymbol{x}_r)' \widehat{\sum}^{-1} (\boldsymbol{x}_q - \boldsymbol{x}_r)}$. To aggregate pairwise similarity, they implement the mean, median, and trimmed mean of the pairwise MDs between the current participant and the participants in each treatment condition: Participants are indexed with treatment condition t using $r \in \{1, ..., R\}$. For each condition t, an average MD between the current participant, q, and the participants previously assigned, t. If the distance in terms of MD for the incoming participant is 2 in the control and 5 for the treatment condition, the incoming participant looks more similar to the control condition. To set the probability of assignment, Moore & Moore (2013) calculate the mean Mahalanobis distances for each incoming participant, q, for all treatment conditions, t, and sort the treatment conditions by these averages. Randomization is biased towards conditions with high scores. For each value of k, with $k \in \{2, 3, ..., 6\}$, the condition with the highest average MD is then assigned a probability k times larger than all other assignment probabilities.

Blocking is thus possible when all covariate information is known at the time of assignment and when this information 'trickles in' over time. Covariate information, however, is only one side of the coin. Researchers also need to take into consideration the characteristics of the variable to block on. Not all types of variables can and should be used the same way to be blocked on. Specifically, the current use of ordinal variables as blocking variables is somewhat problematic.

2.2.4 Ordinal Variables

Ordinal variables are part of the larger framework of categorical variables. Categorical variables represent types of data which are commonly divided into three groups: Nominal, interval, and ordinal variables. Nominal variables are categorical variables with two or more categories that are not intrinsically ordered. Examples include `gender` (Female, Male, Transgender etc.), `race` (African-American, White, Hispanic etc.), and `party ID` (Democrat, Republican, Independent) where the categories cannot be ordered sensibly into highest or lowest. Interval variables are ordered categorical variables with evenly spaced values. Examples include `income` ($20,000, $40,000, $60,000, $80,000 etc.), where the distance between $20,000 and $40,000 is the same as the distance between $60,000 and $80,000. Ordinal variables are ordered categorical variables where the spacing between values is not the same. Examples include `education` (Elementary school, Some high school, High school graduate etc.) where the distance between "Elementary school" and "Some high school" is likely different than the distance between "High school graduate" and "Some college". Each subsequent category has quantitatively more education than the previous, but the exact measure of the distance between the categories is unclear.

For blocking, the categories of nominal variables are often turned into binary variables. This manipulation does not impose any unnatural ordering onto the variable and thus does not require any theoretical assumptions. Interval variables are often made numeric, which is statistically sound. It makes sense to assign numeric values such as 1, 2,

3, and 4 to `income` categories of \$20,000, \$40,000, \$60,000, and \$80,000. The distance between each of these categories is identical between any adjacent pair and thus translates perfectly into the numeric values with equally identical distances. The distance between \$20,000 and \$40,000 is the same as the distance between 1 and 2. Ordinal variables are also often made numeric for blocking. This is problematic because of their unevenly spaced categories. If the `education` categories "Elementary school", "Some high school", and "High school graduate" were turned into the numeric values 1, 2, and 3, we would wrongly assume that the distances between the education categories correspond to these evenly spaced values. Do the numbers 1 to 3 really represent the distances between the categories? Perhaps the true spacing between some of the categories is so narrow they should not even be separate categories at all. We cannot answer this by making an arbitrary assumption that is not justified by the data. Alternatively, if "Elementary school", "Some high school", and "High school graduate" were turned into three separate dummy variables, we would wrongly assume that there is no ordering to these values. In both cases, important information would be lost, which could lead to a large degree of distortion (O'Brien, 1981). To truly use the ordinal nature of a variable, we need to use both its quantitative and its inherent unevenly spaced ordered aspects to make a more underlying description of the data possible (Agresti, 2010). To fill this gap, I borrow from machine learning, which has close connections to problems of causal inference (Grimmer, 2015), and propose an ordered probit model that estimates an ordinal variable's underlying latent continuous structure and is trained on external data.

2.2.5 Ordered Probit Approach

Many approaches in the literature on the analysis of ordinal variables incorporate the distribution of the variable categories (Agresti, 1996). The most promising suggestions focus on natural extensions of probit and logit models (Winship & Mare, 1984) by assigning scores to be estimated from the data (Agresti, 1990) and quantifying each non-

quantitative variable according to the empirical distributions of the variable, assuming the presence of a continuous underlying variable for each ordinal indicator (Lucadamoa & Amenta, 2014). In fact, Agresti (2010) states "that the type of ordinal method used is not that crucial" but that the "results may be quite different, however, from those obtained using methods that treat all the variables as nominal" (p. 3). The same applies to methods which treat ordinal variables as interval (Gertheiss & Tutz, 2008). This suggests that a probit or logit model is suitable to uncover the latent continuous variable underlying an ordinal variable, thus using the ordinal information provided and respecting uneven distances. In the literature, this approach is focused exclusively on the analysis of ordinal variables as a response variable. I propose an ordered probit model that applies to ordinal variables as a predictive variable.

Let there be $\boldsymbol{X}$, an $n \times k$ matrix of explanatory variables. Let further $\boldsymbol{Y}$ be observed on the ordered categories $\boldsymbol{Y}_i \in [1, \ldots, k]$, for $i = 1, \ldots n$, and let $\boldsymbol{Y}$ be assumed to be produced by the unobserved latent continuous variable $\boldsymbol{Y^*}$. $\boldsymbol{Y^*}$ is continuous on $\Re$ from $-\infty$ to ∞. The 'response mechanism' for the r^{th} category is $Y = r \Longleftrightarrow \theta_{r-1} < Y^* < \theta_r$. This requires there to be thresholds on $\Re$: $Y_i^* : \ \theta_0 \underset{c=1}{\longleftrightarrow} \theta_1 \underset{c=2}{\longleftrightarrow} \theta_2 \underset{c=3}{\longleftrightarrow} \theta_3 \ldots \theta_{C-1} \underset{c=C}{\longleftrightarrow} \theta_C$. The vector of (unseen) utilities across individuals in the sample, Y^*, is determined by a linear model of explanatory variables: $\boldsymbol{Y^*} = \boldsymbol{X\beta} + \mathbf{E}$, where $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]$ does not depend on the θ_j and $\mathbf{E} \sim F_{\mathbf{E}}$. For the observed vector $\boldsymbol{Y}$,

$$\begin{aligned} p(\boldsymbol{Y} \leq r|\boldsymbol{X}) &= p(\boldsymbol{Y^*} \leq \theta_r) = p(\boldsymbol{X\beta} + \mathbf{E} \leq \theta_r) \\ &= p(\mathbf{E} \leq \theta_r + \boldsymbol{X\beta}) = F_{\mathbf{E}}(\theta_r + \boldsymbol{X\beta}) \end{aligned}$$

is called the cumulative model because $p(\boldsymbol{Y} \leq \theta_r|\boldsymbol{X}) = p(\boldsymbol{Y} = 1|\boldsymbol{X}) + p(\boldsymbol{Y} = 2|\boldsymbol{X}) + \ldots + p(\boldsymbol{Y} = r|\boldsymbol{X})$. A logistic distributional assumption on the errors produces the ordered logit specification: $F_{\mathbf{E}}(\theta_r - \boldsymbol{X'\beta}) = P(\boldsymbol{Y} \leq r|\boldsymbol{X}) = [1 + \exp(-\theta_r - \boldsymbol{X'\beta})]^{-1}$. The likelihood function is: $L(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y}) = \prod_{i=1}^{n} \prod_{j=1}^{C-1} [\Lambda(\theta_j + \boldsymbol{X}_i'\boldsymbol{\beta}) - \Lambda(\theta_{j-1} + \boldsymbol{X}_i'\boldsymbol{\beta})]^{z_{ij}}$ where $z_{ij} = 1$ if the i^{th} case is in the j^{th} category, and $z_{ij} = 0$ otherwise. The thresholds on $\Re$ partition

the variable into regions corresponding to the ordinal categories. The linear model, Y^*, bins the observations between these thresholds according to the linear predictors.

To use this ordered probit model for blocking, we need to estimate a linear combination of meaningful covariates as predictors and an ordinal variable as the dependent variable. We then train this model on externally and internally valid data. This estimates cutoff thresholds between the ordinal categories and bins data cases according to the linear predictors. The binned cases determine which variable categories make sense, given the underlying latent continuous variable. We then block on the resulting categories.

2.3 Data

Ordinal Probit Model

One of the most common ordinal variables in political science is education. It is widely established that education represents one of the major driving forces behind public opinion and political behavior, such as turnout or donations, in the U.S. (Abramowitz, 2010; Dawood, 2015; Druckman, Peterson, & Slothuus, 2013; Fiorina & Abrams, 2009; Fiorina, Abrams, & Pope, 2011; King, 1997; Leighley & Nagler, 2014). One of the most respected and recognized externally and internally valid data sets are the American National Election Studies. I thus choose the following ordered probit model with the 2016 ANES data (the predictors are standard linear predictors in political science literature):

$$Education \sim Gender + Race + Age + Income + Occupation + PartyID$$

When trained on the 2016 ANES data, this ordinal probit model estimates the thresholds between each of the education categories shown in Table 2.1. The observations in the data are binned according to the estimated threshold coefficients, which in turn determines what education categories make sense, given the underlying latent continuous variable. Figure 2.3 shows the distribution of both the original and the model-estimated education categories. As we can see, all categories 'below' "High school graduate" and

Table 2.1. Ordered Probit Threshold Estimates

Thresholds	Coefficients	Standard Errors	t-values
Up to 1st|1st-4th	−7.869	1.024	−7.681
1st-4th|5th-6th	−7.146	0.717	−9.965
5th-6th|7th-8th	−5.379	0.326	−16.515
7th-8th|9th	−4.671	0.253	−18.472
9th|10th	−3.920	0.206	−19.070
10th|11th	−3.468	0.188	−18.489
11th|12th	−2.984	0.174	−17.100
12th|HS grad	−2.511	0.166	−15.116
HS grad|Some college	−0.710	0.154	−4.607
Some college|Associate	0.384	0.154	2.500
Associate|Bachelor's	1.045	0.154	6.766
Bachelor's|Master's	2.478	0.160	15.538
Master's|Professional	4.099	0.177	23.144
Professional|Doctorate	4.838	0.197	24.589

'above' "Master's" are collapsed because they do not fit the data. The ordered probit model uses the ordinal information with unevenly spaced distances provided and returns categories that do fit the data. We can now use these estimated education categories as the basis for blocking. Assigning numeric values to the new categories is now justifiable because they are based on data-driven estimations. This allows us to block on numerical values with the Mahalanobis distance, which would not be possible without empirical justification. The following sections show that the new estimated categories significantly affect analyses and results.

2.3.1 External Data

We separately block the 2016 ANES on the original and the ordered probit education categories into two treatment groups. We then perform an OLS regression on an interval response variable, a feeling thermometer towards Donald Trump as the Republican presidential candidate. Table 2.2 shows the results. We can see that the differing education categories used for blocking affect the treatment group coefficients, which actually switch

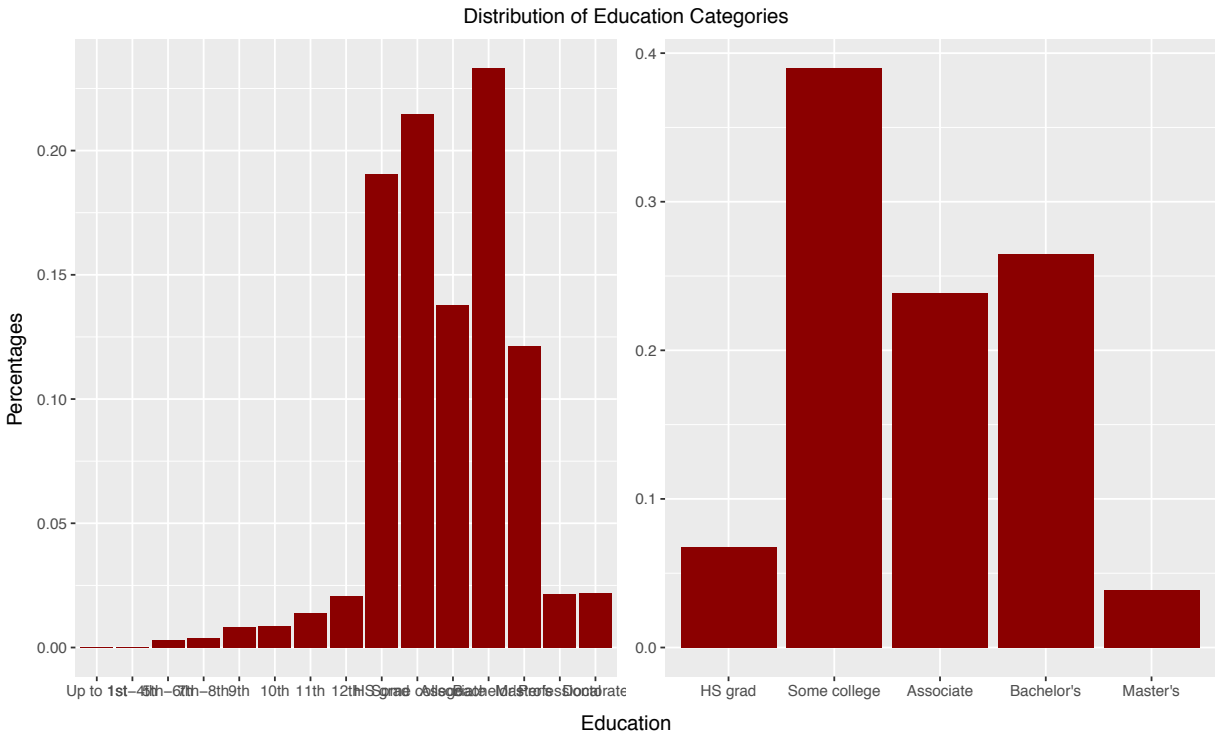

Figure 2.3. Distribution of Education Categories. Original 2016 ANES categories on the left, ordered probit estimated categories on the right

signs (-0.737 v. 1.471).

2.3.2 Original Data

We separately sequentially block participants in an online survey experiment on framing into five treatment groups. The blocking algorithm is performed once on the original ANES edcuation categories for one half of participants and once on the ordered probit education categories for the other half. We then conduct an ordered probit regression on an ordinal response variable on a 5-point Likert scale, ranging from "Strongly oppose" to "Strongly support". As in section 2.3.1, the regression results will show the difference between the two sets of categories.

In order to conduct this experiment, I created an online survey environment based

Table 2.2. OLS After Blocking on Education into 2 Treatment Groups

	Dependent variable:	
	Feeling Thermometer Trump	
	Original Educ Categories	OPM Educ Categories
Group T2	−0.737 (0.971)	1.471 (0.971)
Dem	−22.052 (1.193)	−22.035 (1.193)
Rep	27.364 (1.227)	27.370 (1.227)
Male	4.276 (0.987)	4.268 (0.987)
White	5.249 (2.599)	5.194 (2.598)
Black	−5.878 (3.010)	−5.968 (3.010)
Hisp	−4.174 (2.946)	−4.341 (2.945)
Inc	−1.749 (0.239)	−1.760 (0.239)
Age	0.153 (0.028)	0.153 (0.028)
Constant	30.071 (3.071)	29.066 (3.063)
Observations	3,150	3,150
R^2	0.394	0.394
Adjusted R^2	0.392	0.392
Residual Std. Error (df = 3140)	27.235	27.228
F Statistic (df = 9; 3140)	226.645	226.959

on `R` with `shiny` (Boas & Hidalgo, 2013), as there currently is no available tool to block sequentially online. Popular online survey platforms, such as Qualtrics, do not have offer this functionality, and none of the attempts to combine `R` code work with Qualtrics concern the 'injection' of `R` code into the Qualtrics randomization engine, which blocking would require (Barari et al., 2017; Ginn, 2018; Hainmueller, Hopkins, & Yamamoto, 2014; Testa, 2017). The following is a basic outline of the mechanisms behind this survey environment.

The survey questions, i.e. questions that collect demographic information and questions that apply treatment, need to be designed as `.txt` files and incorporated into a local `shiny` environment. This local environment is then hosted in the cloud and publicly accessible. The hosted website sequentially blocks each incoming participant based on her covariate information and covariate information from all previous participants through constant interaction with the `R` code. The workflow for any incoming participant

is illustrated in Figure 2.4 below:

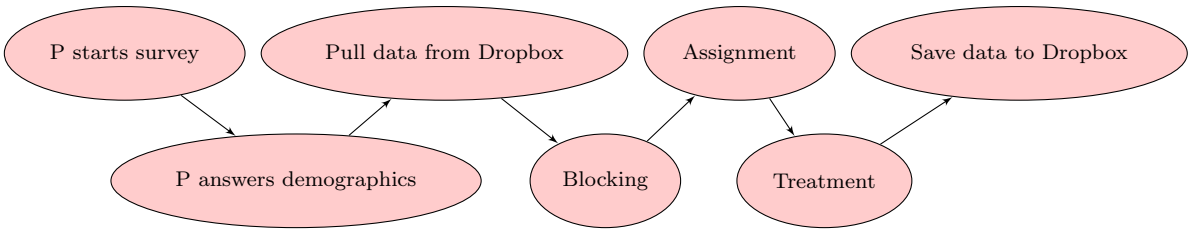

Figure 2.4. Online Survey Experiment Workflow

A participant clicks on the survey link and answers the demographic question. After she selects her level of education, `R` code in the background pulls previous participants' covariate information from a Dropbox server. Based on this information and her chosen education level, the `R` code sequentially blocks and assigns her to a treatment group. The participant then sees and answers the respective treatment question(s). Her responses are then saved on the same Dropbox server. This process is repeated for all incoming participants. If the participant is the first person to take the survey, i.e. if there is no covariate information from previous participants yet, the code randomly assigns her to one of the treatment groups. All subsequent participants are then blocked and assigned as just described.

To recruit participants, the cloud-based website can easily be linked to online market platforms, such as MTurk. MTurk is a service where researchers can host tasks to be completed by anonymous participants. Participants receive financial compensation for their work and Amazon collects a commission. MTurk samples have been shown to be internally valid in survey experiments (Berinsky, Huber, & Lenz, 2012). The use of MTurk in political science experiments has increased dramatically over the past decade and is now common practice (Hauser & Schwarz, 2016). I use Lucid for my experiment, which has been shown to be equally reliable and performs well on a national scale in survey

experiments (Coppock & McClellan, 2019).