PhD Dissertation Prospectus

Advances in Public Opinion Measurement Using Modern Statistical Methods: Sequential Blocking, Mode Effects, and the Power of Moral Arguments

Simon Heuberger
Department of Government
American University

26 July 2018

- Substantive motivation for ordinals (income, education) better. Continue to develop.
- Questions about your proposed measure in Paper II
- Paper III: Respondent definitions of morality. How to measure strength to determine morality → strength?
- MTurk adults vs. AU undergrads: Why would we expect same answers? How to identify "mode effects"?

Abstract

I am writing a methods dissertation that proposes advances in public opinion measurement with modern statistical methods. Particularly, these advances consist of new methods for ordinal survey measures. The dissertation consists of three papers, where two make quantitative methods contributions and one applies both contributions in a substantive analysis in American Politics.

Paper I improves balance in survey experiments by providing a new method of sequential blocking for ordinal covariates. In political science, the most common ordinal covariates are income and education. Both are highly important predictors of public opinion and behavior. It is currently not possible to block on ordinal covariates. I develop a software tool to make this method freely available to other researchers. The tool can be applied to in-person and online survey experiments, as it includes code that provides a finished web-based questionnaire interface. I demonstrate the benefits of my method with simulations, external data, and original data.

Paper II provides a measurement for mode differences in surveys. While there is an abundance of studies that analyze mode differences, existing research can only pinpoint certain parts of surveys where respondents' reactions are and are not affected by mode, the literature offers little help in identifying whether systematic mode effects exist that affect the distributions of responses and the survey instrument as a whole. I develop a measure of entropy that addresses this shortcoming, as traditional statistical approaches used to investigate these differences are not suitable. I test this measure with external and original data and locate its results within the enviroment of the Total Survey Error (TSE).

Paper III uses both methodological contributions from papers I and II and investigates whether moral arguments are part of what makes a political frame strong. First, I conduct a meta-analysis of previous experimental framing studies. Second, I conduct an online poll that tests strong frames from previous experimental studies on their moral content. Third, I conduct a second online poll to gain qualitative insights about how people define moral arguments. Fourth, I design frames with moral and amoral arguments to complement gaps in previous experiments and test these frames in a third online poll. Finally, I design a questionnaire for an online and face-to-face survey experiment that assesses the effect of moral and amoral frames. All polls and the online version of the survey experiment are fielded on MTurk. The face-to-face version of the survey experiment is fielded for an undergraduate population at American University. The sequential blocking method for ordinal covariates developed in paper I is applied when both versions of the experiment are fielded. The ordinal entropy measure developed in paper II is applied in the subsequent analysis of both versions' experimental results. The substantive results show whether moral arguments form a part of what makes a political frame strong.

Paper I: Improving Balance in Survey Experiments with Ordinal Variables

1 Introduction

Survey experiments collect background information and contain questions that represent differing types of treatment, depending on the specific nature of the experiment. Such experiments are created to uncover any effect of a certain treatment on public opinion and behavior. In order to uncover such potential effects, 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. One such option for random assignment is complete randomization. For each participant, the computer flips a coin to decide which treatment group to assign her to (Urdan, 2010). Using complete randomization for large samples results in balance based on the Law of Large Numbers. Using it for small samples, however, can result in serious imbalance. It can easily be that the treatment groups will not look the same. This can leave experimental results in statistically murky waters (Imai, 2018; King et al., 1994; Fox, 2015). In survey experiments, the overall sample size is often split across several treatment groups. Chong and 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. Complete randomization is very unlikely to lead to balanced treatment groups of this size. Researchers need to employ statistical methods to obtain balanced groups here.

One statistical method to ensure balance for small samples is blocking. Blocking involves the arrangement of participants in groups that are similar to one another in terms of the participants' covariates, i.e. their background information. Within these groups, participants are then randomly assigned. This is different from complete randomization, where partic-

ipants are immediately randomly assigned without taking their covariate information into account.

To apply blocking, a researcher needs to know all the covariate information for all participants before assigning treatment. Oftentimes, however, this is not possible as participants arrive for assignment at differing times. The researcher knows the covariate information of the arriving participant and (if she is not the first one) of the previous already assigned participants, but she does not have any information about future incoming participants. She thus needs more advanced statistical methods to block these participants. One such method is sequential blocking. Sequential blocking assigns each participant based on information from previously assigned participants and the incoming participant herself, while participants are still arriving to be assigned. Sequential blocking uses statistical methods to 'decide' which group to assign a participant to. These measurements differ depending on the form that the covariate information comes in. Some covariate information is binary (e.g. gender) while other is categorical (e.g. party ID, race), to name only two. Research has shown that sequential blocking greatly improves balance in small samples (Moore and Moore, 2013).

While sequential blocking is possible for a variety of variable forms, there is currently no method to sequentially block on ordinal variables, such as income and education. Substantively, this means that a researcher currently cannot block incoming participants based on their provided level of education and income. This is problematic for political science, as education and income represent two of the strongest predictors of political behavior. I develop a new method of sequential blocking for ordinal variables that fills this gap. This method is freely available to other researchers as a package using the open-source software R. The R package provides functions that allow researchers to apply sequential blocking for ordinal variables in both in-person and online survey experiments. Substantively, I employ this method in simulated data, external data from political science experiments published in peer-reviewed journal articles, and original data in the form of an online and face-to-face

survey experiment on the importance of moral arguments in political framing. The following sections provide background on sequential blocking and complete randomization, showcase the statistical development of the new ordinal sequential blocking method, and give more detail about the data to be used with this method.

2 Theory

2.1 Preliminary Notations on Randomization

The simplest of experiments has two potential outcomes for participants i: y_{1i} and y_{0i}, with 1 denoting the treatment and 0 referring to the control. If a researcher, for example, intends to analyze the effect of a 2-week mathematics training camp on high school students' performance in exams, she devises an experiment where one half of her student sample participates in the camp (treatment) and the other half abstains (control). At the end of the camp, both groups take the same exam. If both groups of students look the same regarding their covariates (age, mathematical skill, intelligence etc.), a comparison of the groups' average test results reveals the Average Treatment Effect (ATE):

$$\beta_i = y_{1i} - y_{0i}$$

More specifically, for a sample of n students, this comparison reveals the Sample Average Treatment Effect (SATE):

$$\frac{1}{n}\sum_{i=1}^{n}(y_{1i} - y_{0i})$$

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 well student A performs in the exam after participating in the camp whilst also observing how well the same student A would have performed without taking part in the camp. If we could, it would be simple to calculate the True Average Treatment Effect (TATE) (Moore, 2012):

$$\overline{TE} = \frac{1}{2n}\sum_{i=1}^{2n}(y_{1i} - y_{0i})$$

Since the TATE 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 students look the same in every measure, we can use the students who participated in the camp (treatment) to estimate what would have happened to the students who did not participate (control) if the abstaining students had in fact attended the camp. The crucial aspect is of course whether the two groups do indeed look the same in terms of the students' covariates. There are two main mechanisms by which this can be achieved: Complete randomization and blocking.

2.2 Complete Randomization and Blocking

Complete 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). Complete randomization increases covariate balance as n increases (Imai et al., 2009). The larger a researcher's sample, the better the resulting balance from complete randomization. Complete randomization enables the comparison of the SATE to be free from systematic error, which allows the researcher to attribute any treatment effects to the treatment (King et al., 2007).

While complete 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 complete randomization will be balanced. With small samples, however, it is possible to get 'unlucky' and end up with unbalanced groups (Imai et al., 2008). Blocking can help achieve balance in such scenarios (Epstein and King, 2002).

Blocking involves the arrangement of participants in groups that are similar to one an-

other in terms of the participants' covariates. As mentioned above, the key aspect in experimental studies is whether the treatment and control groups look the same. In complete randomization, this is achieved by random chance. In blocking, this is achieved by using covariate information about the participants. Researchers use observed covariates to create similar treatment and control groups before treatment is assigned. Blocking is thus better suited to achieving balance in finite samples, as it "directly controls the estimation error due to differing levels of observed covariates in the treatment and control groups" (Moore, 2012, p. 463).

Efficient blocking focuses on covariates that affect the outcome. In U.S. politics, for instance, it is widely established that party ID represents one of the major driving forces behind public opinion (Abramowitz, 2010; Druckman et al., 2013; Fiorina et al., 2011; King, 1997). Making sure that participants in treatment and control groups have identical levels of party ID is thus crucial in experiments that test the effect of interventions on public opinion.

2.3 Blocking v. Post-Experiment OLS Controls

Researchers often employ complete randomization in survey experiments and control for covariates post-hoc in OLS regression models. This is done to increase balance and model precision. Proponents of this approach might question the need for blocking approaches if post-hoc OLS controls can achieve similar balance and are much less computationally and mathematically complex. While controlling for covariates undoubtedly improves balance over 'basic' complete randomization, it does not achieve the level of balance that blocking can provide. Most crucially, this is because any OLS model requires assumptions about the choice of controls to include. More often than not, it is very difficult if not impossible to determine all needed controls. Such assumptions are not required in blocked experiments. Additionally, blocking balances variation, which post-hoc controls cannot provide. Finally, it stands to reason that ex-ante balance prior to treatment is preferable to post-hoc adjustments as it removes the possibility for errors to be introduced.

2.4 *Sequential Blocking*

Sequential blocking is a special form of blocking. Generally, there are two main starting positions for the researcher who wants to conduct an experiment: (1) The researcher knows all covariate information of all participants at the time of randomization, and (2) covariate information of participants 'trickles in' as the experiment progresses and participants arrive for assignment. Complete randomization, i.e. flipping a coin for each participant to be assigned to treatment or control, is possible in both positions. 'Normal', or for our purposes 'nonsequential', blocking can only be used in scenario 1: The researcher possesses all information about all participants when the experiment begins. Sequential blocking, on the other hand, is applied when the researcher does not know all covariate information of all participants when the experiment begins. Instead, she only possesses the covariate information of the current incoming participant and previously assigned participants, but not future incoming participants. This is scenario 2. Sequential blocking is thus blocking 'on the go'.

When all participant characteristics are known when the experiment begins (scenario 1), sequential blocking is not necessary, as no participants arrive individually for treatment. In many political experiments, researchers have an already-collected data set in front of them at the start of the experiment and then randomize a set of households, precincts, or individuals to treatments all at once. All covariate information on all participants is known, i.e. the characteristics of all participants are known at the time of randomization. Examples are any studies that use pre-existing databases. A prominent example are the American National Election Studies (ANES) that are often used to analyze voter turnout (see for instance Jackman and Spahn, 2018; Leighley and Nagler, 2014).

In these experiments, 'nonsequential' blocking suffices to create homogeneous groups within which treatments can be assigned. In an experiment of scenario 2, however, the researcher has much less information about the eventual full sample to use in assigning treatment than in an experiment with pre-existing data sets. More advanced statistical methods are thus required to exploit available background information to block 'on the

go', i.e. to block sequentially. Chow and Chang (2007) identify four types of sequential experiments, which are shown in Table 1.

Table 1: Different Randomization Designs For Sequential Experiments

Non-Adaptive	Assignment probabilities (AP) fixed throughout the trial
Treatment-Adaptive	AP change based on numbers of participants already in treatment groups
Response-Adaptive	AP change as a function of previous participants' outcomes
Covariate-Adaptive	AP change based on previous participants and the current participant

Adapted from Chow and Chang (2007)

Non-adaptive randomization and treatment-adaptive randomization can (and often do) ignore the researcher's detailed data on participants. They thus leave important information lying on the table. Response-adaptive randomization varies probabilities based on knowledge about previous participants' outcomes. This is likely very useful for clinical trials, where one outcome is often the clear goal: Consider a test where the treatment group receives an experimental medication to fight cancer and the control group receives a placebo. If it becomes clear as the trial progresses that the medication is effective in fighting cancer, it makes sense to assign more people to treatment, as this could potentially save their lives. In political science, this is not the case. Assessing whether a certain survey treatment affects participants' support for a political issue is not a matter of life and death. This leaves covariate-adaptive randomization (CAR), which varies probabilities based on knowledge about previous participants and the current participant, which appears ideally suited.

2.4.1 Basic Covariate-Adaptive Randomization (CAR) Approaches

There are two traditional approaches to covariate-adaptive randomization (CAR). The first approach is the biased coin design developed by Efron (1971), which sets the current participant's treatment assignment probability using its entire covariate profile at once. I follow Chow and Chang's brief and excellent description of Efron's design (Chow and Chang, 2007) here:

$$P(\delta_j|\Delta_{j-1}) = \begin{cases} 0.5 & \text{if } N_A(j) = N_B(j), \\ p & \text{if } N_A(j) < N_B(j), \\ 1-p & \text{if } N_A(j) > N_B(j). \end{cases}$$

δ_j is a binary indicator for treatment assignment of the jth participant ($\delta_j = 1$ for treatment A; $\delta_j = 0$ for treatment B). $\Delta_{j-1} = \{\delta_1, ..., \delta_{j-1}\}$ is the the cumulative set of treatment assignments up to participant $j-1$. Imbalance is measured by

$$|D_n| = |N_A(n) - n|.$$

The second approach is minimization, developed by Pocock and Simon (1975), which considers covariates one at a time and can limit marginal imbalance across an arbitrarily large set of covariates. I follow Rosenberger and Lachin's overview of the method here (Rosenberger and Lachin, 2002): Let $N_{ijk}(n), i = 1, ..., I, j = 1, ..., n_i, k = 1, 2$ (1 = treatment A, 2 = treatment B), be the number of participants in category j of covariate i on treatment k after n participants have been randomized. Suppose the $(n+1)$th participant to be randomized is a member of categories $r_1, ..., r_I$ of covariates $1, ..., I$. Let $D_i(n) = N_{ir_i1}(n) - N_{ir_i2}(n)$, with the weighted difference measure being $D(n) = \sum_{i=1}^{I} w_i D_i(n)$. w_i are weights the researcher chooses based in the perceived importance of covariates. If $D(n)$ is lower than 0, then the weighted difference measure indicates that treatment B was chosen more often so far for that set, $r_1, ..., r_I$, of strata, and that participant $n+1$ should be assigned with higher probability to treatment A. This argument holds vice versa if $D(n)$ is greater than 0. According to Pocock and Simon, the coin should be biased with

$$p = \frac{c^* + 1}{3}$$

and the following rules implemented: If $D(n) < 0$, assign the next incoming participant to treatment A with probability p. If $D(n) > 0$, assign the next incoming participant to treatment A with probability $1-p$. Finally, if $D(n) = 0$, assign the next incoming participant to treatment A with probability $1/2$, where $c^* \in [1/2, 1]$.

The biased coin and minimization approaches use discrete covariates with a very small

number of levels (Moore and Moore, 2013). Discrete covariates allow for simpler pairing procedures, as the number of possible covariate levels is finite. This is not the case for continuous covariates, where the number of possible covariate levels rises exponentially. Blocking on continuous covariates is thus not possible with these traditional approaches (Markaryan and Rosenberger, 2010; Rosenberger and Lachin, 2002; Eisele, 1995).

2.4.2 Extended CAR for Continuous Covariates

In the standard approaches, the incoming participant is assigned to the treatment group with the fewest participants with identical covariate information. With continuous covariates, no two participants will look the same, and thus identical participants do not exist. Moore and Moore (2013) develop a method to include continuous variables in sequential blocking with CAR by exploiting relationships between the current participant's covariate profile and those of all previously assigned participants.

Moore and Moore (2013) 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 and Moore (2013) test several methods, most notably a set of strategies that calculates the mean Mahalanobis distances for each incoming participant, q, for all treatment conditions, t, and sorts 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, $T - 1$.

2.4.3 Extended CAR for Ordinal Covariates

Moore and Moore (2013) extend basic blocking approaches to provide a method to apply sequential blocking to continuous covariates. They show that their approach significantly improves balance for some experiments. I extend blocking approaches further by developing a method to sequentially block on ordinal covariates. My method also extends blocking methodology to previously unused multi-category treatments.

Consider the following simple example for a common ordinal variable in political science: Income. Income is a strong predictor for political behavior such as turnout, representation, or donations (Dawood, 2015; Fiorina and Abrams, 2009; Leighley and Nagler, 2014). A researcher conducts a basic experiment with one treatment and one control group. The researcher also collects information about income on three levels: Low, medium, and high. For all already-assigned participants, let the income distributions across these three levels look like figure 1 for both groups.

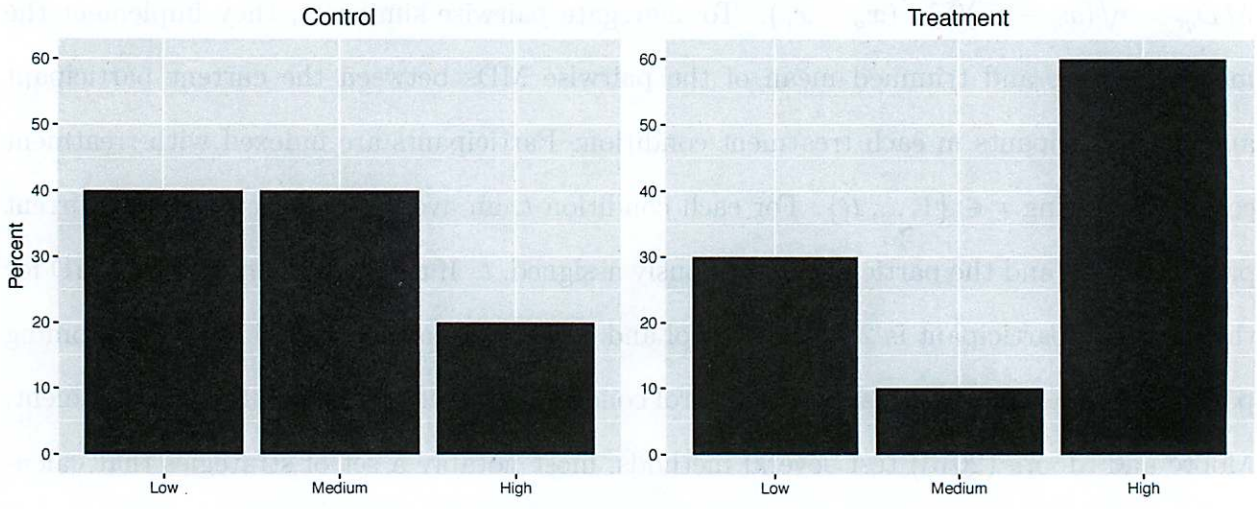

Figure 1: Example Distributions of 3-Level Income in Treatment and Control Groups

A new participant arrives with low income. We need a suitable measure, or measures, to define the dissimilarity between this incoming participant and the already assigned par-

ticipants. It would be easy to transform these income levels into the numeric values of 1, 2, and 3, and simply apply the Mahalanobis distance used for continuous variables. In doing so, however, we would assume that the distance between 1 and 2 is the same as the distance between 2 and 3. We do not have to make such a strong assumption if we keep the ordinal levels low, medium, and high. Similarly, we could transform low, medium, and high income into three dummy variables. Income would then look like non-ordinal variables such as race or gender. Again, crucial information would be lost. We cannot employ the mean because there is no mean for ordinal variables. Exact matching, and with that Coarsened Exact Matching (Iacus et al., 2011), might suggest itself, since income is not a continuous variable. Unlike matching, however, we are not looking to create identical pairs of participants. We are comparing one participant with one specific level of income with the whole distribution of income across many other, previously-assigned participants. The distribution rarely has one specific level, therefore there will not be exact matches for the majority of incoming participants. In order to exploit the unique information contained in ordinal variables, we need different measures.

I propose a weighted approach that incorporates all levels of the respective ordinal variable in question. Using only the incoming participant's income level of the distribution for comparison would ignore all the available information on the two other income levels, medium and high, in all treatment groups, and treat income as a one-level variable, thereby removing its very nature of an ordinal variable. I propose to develop the following algorithm: Let an ordinal variable have levels $m_i = m_1, m_2, ..., m_I$. Let the level of an incoming participant be m_1. Let the number of treatment groups be $t_j = t_1, t_2, ..., t_J$, where J can take on any number to account for binary and multi-category treatments. For each treatment group, we calculate the proportion of participants with level m_1, $p(m_1)$, and weight this proportion by $\frac{I}{J+1}$. For each treatment group, we then calculate the proportions of participants in the next level m_2, $p(m_2)$, and weight this proportion by $\frac{I-1}{J+1}$. This is continued until the proportion of participants for the last level m_I, $p(m_I)$, which is weighted by $\frac{1}{J+1}$. This gives us S_j, the

weighted average of the income distribution for each treatment group:

$$S_j = \frac{I}{J+1} \times p(m_1) + \frac{I-1}{J+1} \times p(m_2) + \frac{I-2}{J+1} \times p(m_3) + ... + \frac{1}{J+1} \times p(m_I)$$

Note that the proportions over all the treatment groups sum up to one, $\sum_{i=1}^{I} p(m_I) = 1$. Note also that the sum of all weights equals $\frac{I}{2} \times (I+1)$. The numerators for the respective level proportions are set in descending order because m_1, as the ordinal variable level of the incoming participant, is the most important in the distribution, followed by m_2, then m_3, and so on. Since we are adding up weighted proportions, a higher weighted average S_j represents more similarity to the incoming participant. She will thus be assigned to the treatment group with the lowest S_j.

In the simple income example given above, the income variable has three levels, thus $I = 3$ (Low, Medium, and High Income), and there are two treatment groups, thus $J = 2$. The incoming participant has low income, thus $m_1 = \text{Low}$. In the control group, $p(\text{Low}) = 0.4$, $p(\text{Medium}) = 0.4$, and $p(\text{High}) = 0.2$. S_1 for the control group ~~thus~~ is:

$$S_{\text{Control}} = \frac{3}{2+1} \times 0.4 + \frac{2}{2+1} \times 0.4 + \frac{1}{2+1} \times 0.2 = 0.733$$

In the treatment group, $p(\text{Low}) = 0.3$, $p(\text{Medium}) = 0.1$, and $p(\text{High}) = 0.6$. S_2 for the treatment group thus is:

$$S_{\text{Treatment}} = \frac{3}{2+1} \times 0.3 + \frac{2}{2+1} \times 0.1 + \frac{1}{2+1} \times 0.6 = 0.567$$

We thus assign the incoming participant with low income to the treatment group.

2.5 *Computational Development*

2.5.1 R Package: BlockExperiments

The new sequential blocking method I develop is built in R and made freely available to other researchers as the R package `BlockExperiments`. `BlockExperiments` includes the functions provided by `blockTools` (Moore and Schnakenberg, 2016) and further improves on the state of the art. By incorporating `blockTools`, `BlockExperiments` thus allows the user to apply sequential blocking to binary, continuous, and ordinal variables. It also includes web-components that allow the user to create a finished web-based survey questionnaire that can be directly fielded online. `BlockExperiments` can thus be employed for in-person sequential survey experiments as well as online sequential survey experiments.

Figure 2 shows a sample workflow for the use of `BlockExperiments` in in-person experiments. Processes performed by the software in R are shown as blue boxes. Actions that need to be completed by the user outside of R are shown as red clouds.

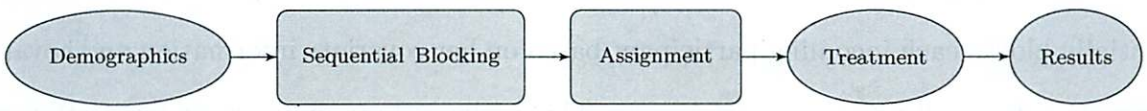

Figure 2: In-Person Survey Experiment Workflow

The user creates the survey questionnaire that collects demographic information from each incoming participant. This information is uploaded into R. `BlockExperiments` uses this information and information from all previous participants to sequentially block the incoming participant and then assign her to a treatment group. The participant then receives the treatment designed by the researcher, and her responses are saved. This process is repeated for all incoming participants.

There is currently no way to use sequential blocking when surveys are fielded online. `BlockExperiments` provides a simple interface to do so, as shown in Figure 3. As before, processes performed by the software in R are shown as blue boxes, while user actions that

need to be completed outside of R are shown as red clouds.

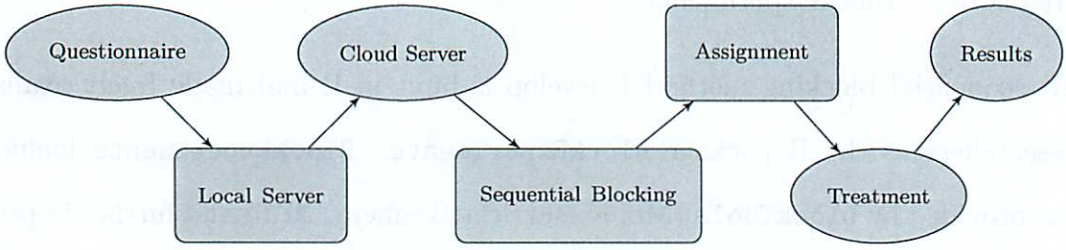

Figure 3: Online Survey Experiment Workflow

The user uploads the complete survey questionnaire, i.e. questions that collect demographic information and questions that apply treatment, as a `.csv` file into R. The `shiny` function within the R package creates a web survey structure by employing the user's R environment as the local 'host server'. This means that the survey is not accessible on the web yet – it is located on the user's local computer environment. To make the survey public, the user hosts this website on a cloud server (e.g. Amazon Web Services, Blue Ocean). The created cloud-based public website is used directly to recruit participants. The website sequentially blocks each incoming participant based on her covariate information and covariate information from all previous participants. This is done through constant interaction with the R code provided by `BlockExperiments`, which is integral to the website build. Based on the sequential blocking, the incoming participant is assigned to a treatment group and then receives the appropriate treatment. Her responses are then saved on the server. This process is repeated for all incoming participants.

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 et al., 2012). The use of MTurk in political science experiments has increased dramatically over the past decade and is now common practice (Hauser and Schwarz, 2016).

For online survey experiments, `BlockExperiments` thus creates its own web-based survey. A theoretic alternative to this approach is to use online survey design platforms, such as Qualtrics. Since Qualtrics is very popular, it would appear fruitful to have `BlockExperiments` load questions directly into Qualtrics, instead of hosting the questions on a remote website via `shiny`. However, this idea appears not computationally feasible. There have been attempts to combine R code work with Qualtrics:

Hainmueller et al. (2014) show how conjoint analysis enables researchers to estimate the causal effects of multiple treatment components and assess several causal hypotheses simultaneously. To do so, they develop a computer program that creates conjoint question design templates. The program exports the templates as a `.php` file, which is then uploaded to a web server, which can then in turn be loaded into Qualtrics via the Qualtrics functions Web Service and Embedded Data. This pipes the file into the Qualtrics question design environment. While this facilitates conjoint survey questionnaire design, it does not affect Qualtrics's randomization engine but 'only' loads question categories. Barari et al. (2017) develop the R package `cjoint`, which is designed to calculate Average Marginal Component Effects of conjoint survey data. It has a function called `makeDesign` which creates the conjoint survey design from output created by Hainmueller et al. (2014)'s Conjoint Survey Design Tool and also contains two functions that pull the final `.csv` file from Qualtrics directly into R. This tool then addresses the data analysis after the survey has run, but not Qualtrics's randomization engine. Similarly, the R package `QualtricsTools` (Testa, 2017) provides functions to analyze Qualtrics survey data by automatically processing the data into reports breaking down the responses to each question. A similar function is offered by the R package `qualtRics` (Ginn, 2018), which pulls survey data from Qualtrics to analyze directly in R through the Qualtrics APIs.

None of these tools, then, concern the 'injection' of R code into the Qualtrics randomization engine. Boas and Hidalgo (2013) recommend using the R web tool `shiny` to achieve this as it is purposefully built to easily harnesses all the statistical power from R. It thus

seems a much more feasible vehicle for my project.

3 Data

I demonstrate the improvements produced by my sequential blocking method by comparing balances in data with ordinal covariates from several data sources: Simulations, others' external data, and my own original survey data. In each source of data, I focus on the two ordinal variables that are most common and most important in political science: Income and education.

3.1 Simulated Data

I simulate an experiment with common demographic covariates, such as age, race, gender, income, education, and party ID. The sample size is 1,000 participants, which is a common sample size for political science survey experiments. The experiment features 5 treatment groups, which results in around 200 participants per treatment group. Compared to prominent survey experiments such as Chong and Druckman (2007), who have fewer than 900 participants for 17 treatment groups, this is a conservative setup. The outcome variable is irrelevant, since we are only interested in balance here. I employ two separate randomization methods: Complete randomization, and my sequential blocking method. Each method is simulated 1,000 times. This means I simulate 1,000 runs of this experiment with complete randomization, and another 1,000 runs of this experiment with my sequential blocking method. The comparison of the distribution of the blocked d^2 p-values and the complete randomization balance demonstrates balance improvements produced by my method.

3.2 External Data

I also use data from external experimental survey studies and sequentially block the participants in these data for ordinal covariates. To create a sequential nature, I assume that the order of the observations in the replication data file represents their entry order into a

sequential experiment. One example for such data is Tomz and Weeks (2013), who explore whether American participants are more likely to support pre-emptive military strikes on non-democracies versus democracies. The authors present participants with different country profiles and ask participants whether they would support pre-emptive American military strikes against the hypothetical country. They randomly assign various characteristics of these profiles, including (1) whether the country is a democracy, (2) whether the country has a military alliance with the United States, and (3) whether the country has a high level of trade with the United States. As before, however, the outcome variable is irrelevant, since we are only interested in balance. I plan to use up to four suitable similar experimental survey data here, including Hainmueller et al. (2014)'s conjoint experiment data. If possible, I also plan to obtain the data from Chong and Druckman (2007) (which is currently not publicly available).

With published peer-reviewed data, it is to be expected that the experiments are overall well-balanced, i.e. $p \approx 0.9$ (Moore and Moore, 2013). Nonetheless, I expect a comparison of the distribution of the blocked d^2 p-values and the balance in the original experiment to improve balance. Moore and Moore (2013) show that sequential blocking can indeed make overall balance in well-balanced experiments nearly perfect.

3.3 Original Data

For the final application, I employ original survey data. I design a questionnaire to assess the effect of moral frames on public opinion in a survey experiment. This questionnaire is fielded online on MTurk for a random sample of U.S. adults and face-to-face for undergraduate students at American University. In each experiment, I use my sequential blocking method on ordinal covariates to assign participants to treatment groups. While I analyze the substantive results in depth in paper III, they are irrelevant here, as we are only interested in balance. I then use the collected covariate information and re-assign participants to treatment groups with complete randomization. As before, the comparison of the distributions of the blocked

d^2 p-values and the complete randomization balance demonstrates balance improvements produced by my method.

Paper II: Measuring Mode Effects for Ordinal Survey Responses in Online and Face-to-Face Survey Experiments

1 Introduction

Survey methodology is continuously undergoing changes. One prominent area of such change is survey mode. Survey mode refers to the manner in which survey responses are collected. The most common survey modes are face-to-face interviews, mailed letters, telephone calls, and online questionnaires. Up until the 1970s, mail and face-to-face surveys were the main modes of data collection (Lyberg and Kasprzyk, 1991). Because of the low costs of telephone calls and based on the ever-increasing phone network coverage across the US, survey researchers then gradually shifted to telephone calls, which soon replaced face-to-face interviews as the most widely used survey mode, particularly through the use of random-digit dialing (Dillman, 2000). It is much cheaper to call potential respondents from your desk than to travel to each respondent's place of residence for an interview. In the 1990s, the emergence of the internet challenged the supremacy of the telephone to conduct surveys (Brick, 2011). It is much cheaper still to create one digital questionnaire and provide it online than to call and talk to potential respondents for hours.

Changes in mode choice can be problematic. Each mode can influence the way participants think about and respond to survey questions, which in turn hurts comparability of surveys across modes. This influence is called 'mode effects'. Mode effects form a key part of the Total Survey Error (TSE) framework. The TSE consists of five primary sources of error in survey research: Sampling error, coverage error, nonresponse error, measurement error, and post-survey error (Weisberg, 2005). It comprises all potential errors that can occur whilst conducting a sample survey and using model-based statistical means to describe a population (Biemer, 2010a). Survey mode choice can 'activate' each and all of these sources of error (Groves and Lyberg, 2010; Weisberg, 2005; Ansolabehere and Schaffner, 2014; Atkeson et al., 2014; Ye et al., 2011; Yeager et al., 2011; Bowyer and Rogowski, 2017).

While research is able to pinpoint certain parts of surveys where respondents' reactions are and are not affected by mode, the literature offers little help in identifying whether systematic mode effects exist that affect the distributions of responses and the survey instrument as a whole. For instance, is there a measurement error in the form of a critical distributional difference in aggregate survey responses? If there is such a distributional measurement error, what potential do such a measurement and its related findings have? Can it affect how we interpret question responses given on each respective mode? Do such systematic differences occur, and if so, when do they occur, and do they matter?

In an attempt to address this question, Homola et al. (2016) develop an entropy measurement (Shannon, 1948) for discrete survey measures. Entropy roughly translates to 'information' and is most often used in information theory to carry levels of information in a message. Entropy represents a measure of the average amount of information that is required to describe the distribution of a variable of interest (Cover and Thomas, 1991). They find that entropy catches differences in mode effects that traditional statistical measurements that assume continuous data, e.g. the standard deviation, miss: Mode effects do not show in measures of centrality, but instead are shown by greater variability. Homola et al. (2016) show that entropy is a good measure of variability for discrete survey measures as it reveals differences in how respondents react to identical questions presented in two differing modes.

Entropy thus shows great potential to uncover distributional differences in responses between survey modes. In its application to political science data, however, Homola et al.'s measure falls short on the aspect of ordinality. In political science, two of the most important predictors of political behavior are ordinal survey measures: Education and income. Income, for instance, is a strong predictor for political behavior such as turnout, representation, or donations (Dawood, 2015; Fiorina and Abrams, 2009; Leighley and Nagler, 2014). Homola et al. (2016) do not account for ordinality in their model and there is currently no way to measure entropy for these important ordinal survey measures. I fill this gap by developing such a measurement. I apply it to two modes: Online and face-to-face. Substantively, I

employ this measure in external data from political science experiments published in peer-reviewed journal articles and original data in the form of an online and face-to-face survey experiment on the importance of moral arguments in political framing.

The following sections describe the TSE framework in more detail and ~~put it in~~ connection with research on mode effects, showcase Homola et al.'s entropy measurement, develop a measurement for ordinal entropy, and provide a brief overview of the data sources I use to employ this measurement.

2 Theory

2.1 Total Survey Error (TSE)

~~As mentioned above~~, the TSE is an overarching framework that gives a complete account of the potential errors that can result from conducting a sample survey and using model-based statistical means to describe a population (Biemer, 2010a). It "is based on analyzing the several different sources of error in surveys and considering how to minimize them in the context of such practical constraints as available money" (Weisberg, 2005, p. 16). 'Error' here refers to the difference between the values the researcher obtains and the 'true value' she should obtain.

There are five primary sources of error in the TSE: Sampling error, coverage error, non-response error, measurement error, and post-survey error (Groves and Lyberg, 2010; Groves et al., 2009; Weisberg, 2005). Survey mode can affect each and all of these sources. They are diagrammed in Figure 4 which shows the order of concern that each potential source of error appears in the survey data collection process. The arrows in the figure represent this sequential nature, rather than causal paths. Figure 5 locates the five potential sources of error more precisely within the survey life cycle.

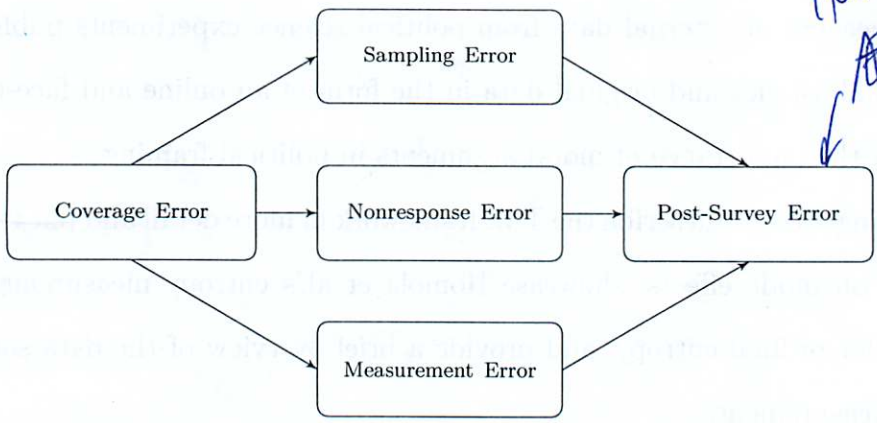

Figure 4: Model of Total Survey Error (from Homola et al. 2016)

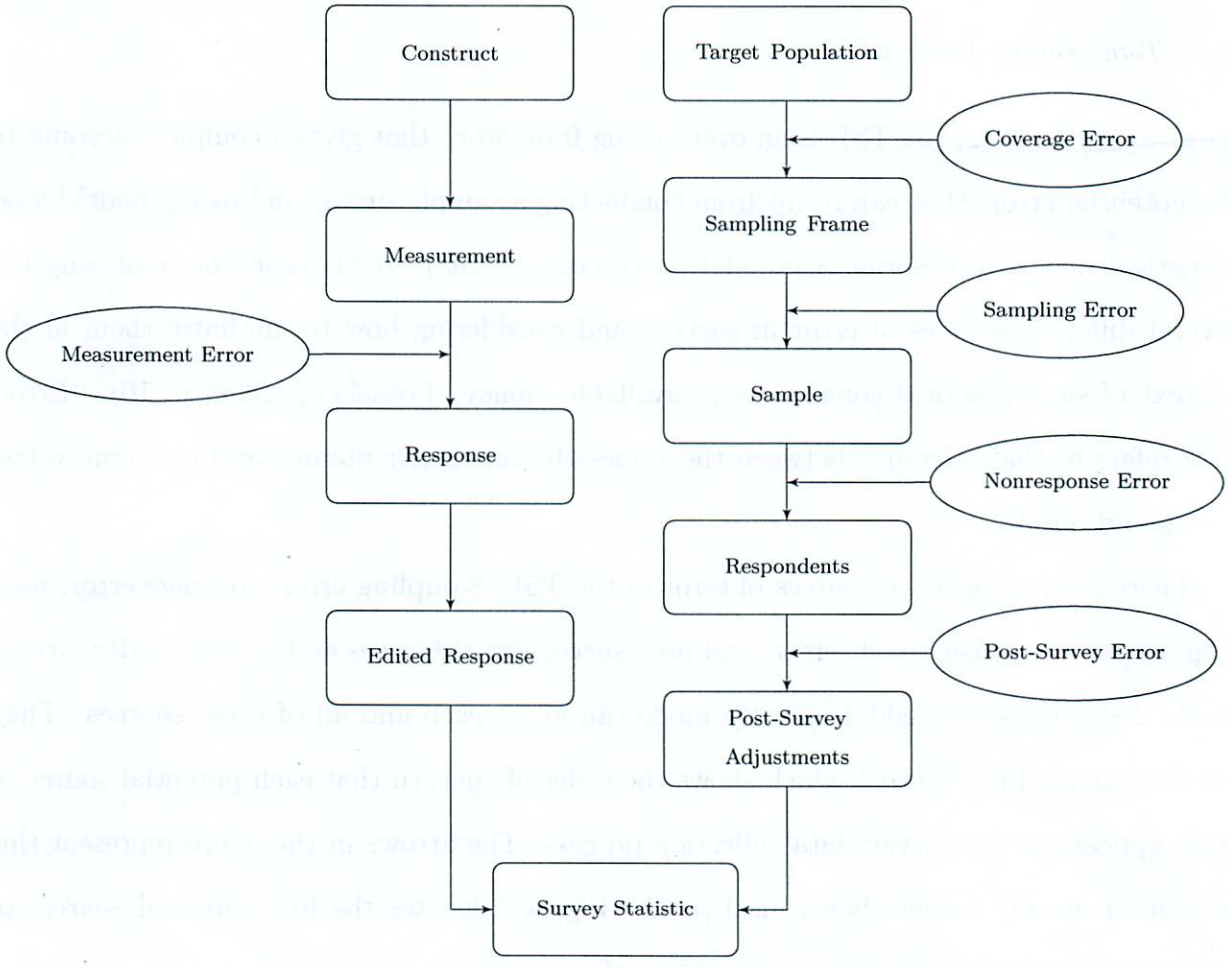

Figure 5: Survey Life Cycle (adapted from Groves et al. 2009)

As shown in Figure 5, all sources add up to form one resulting total survey error, or TSE. I now address each of the five sources and their explicit connection to survey mode in turn.

2.2 Five Types of Survey Error in Connection with Survey Mode

2.2.1 Coverage Error

Coverage error applies to the nonobservational gap between the target population and the sampling frame, i.e. the actual set of units from which the sample is taken (Groves et al., 2009). This means we choose a sample intended to represent the target population from a sampling frame that does not actually cover that target population. In statistical terms, coverage error is the mathematical difference between a statistic calculated for the population that is studied and the same statistic calculated for the target population (Weisberg, 2005).

In early survey research days, potential survey respondents could only be contacted through mailed letters. Since everyone in the target population, i.e. the U.S. population, had a mailing address, the sampling frame of mailing addresses was identical to the target population, both for questionnaires sent directly through the mail and for mailed requests for face-to-face interviews. The invention of the telephone provided a cheaper, faster alternative. Until around 35 years ago, however, telephone landline access did not cover the entire U.S. population, so any samples drawn from a list of landline phone numbers did not cover the entire adult U.S. population. In such a case, there are population units that fall outside of the sampling frame because of the chosen vehicle of communication, i.e. the survey mode. People who did not have telephones could not be contacted and were thus not part of the sampling frame. Since people who did not yet have telephones were likely to differ systematically from people who did have them, this lead to biased estimates of the variables of interest. In other words, bias occurred because non-covered units were omitted from the sample because of the choice of survey mode.

In more ~~current~~ recent times, this issue arises for people who no longer have a landline but instead rely exclusively on their cell phone. If the chosen vehicle of communication is landline telephone numbers, these people will be excluded from the sample. In 2008, 17.5 percent of U.S. households had only wireless telephone access, with landline coverage having decreased

to 80 percent (Blumberg et al., 2008), with this trend likely to have continued in the years to follow. Similarly, coverage error can be substantial in online surveys intended to represent the U.S. population, as 11 percent of U.S. adults currently do not have access to the internet (PewResearchCenter, 2018). This number continues to shrink, however, with the continuous spread of high-speed internet access in rural areas.

2.2.2 Sampling Error

Researchers can only very rarely ask an entire population but instead have to rely on samples. As a result, not all people in the sampling frame, i.e. the entire population, are part of the eventual sample. This nonobservational gap between the sampling frame and the sample is called sampling error (Groves et al., 2009). Sampling error is thus very much by definition part of survey data collection. It is often colloquially known as the 'margin of error', i.e. the difference in estimates between repeated samples.

Sampling error primarily arises in the form of sampling bias. Sampling bias occurs when some units of the sampling frame are systematically excluded from the sample. This can be avoided by probability sampling, i.e. randomly selecting units of the sampling frame to be included in the sample. It is essential here that everyone in the larger population has a known probability of being chosen for the random sample. This is not the case with non-probability sampling, also known as convenience sampling, where researchers collect responses of everybody who conveniently 'stops by' to answer questions. In non-probability sampling, sampling bias can be systematic. While this can affect all survey modes, it is a particular problem for online surveys, where participants volunteer to answer questions, often in return for financial compensation (Ansolabehere and Schaffner, 2014). One such example is MTurk (see section 2.5.1 of Paper I for a brief outline of MTurk). These volunteers can differ systematically from participants who are randomly selected and can look very different in terms of their covariates than the rest of the population (Dillman et al., 2014; Couper, 2000; Hays et al., 2015; Alvarez et al., 2003).

2.2.3 Measurement Error

Measurement error is "the error that occurs when the measure obtained is not an accurate measure of what was to be measured" (Weisberg, 2005, p. 18). In other words, the questions that researchers choose to measure something do not actually measure that thing. Measurement error is the observational gap between the ideal measurement and the response obtained (Groves et al., 2009). There are two primary levels of measurement error: Respondent-induced measurement error and interviewer-induced measurement error. Respondent-induced measurement error occurs when respondents do not give the answers they 'should' give. The most common respondent-induced measurement errors are non-differentiation and speeding. IN WHAT KINDS OF SURVEYS?

Non-differentiation involves participants who give answers that simply enable them to finish the survey as quickly as possible. For online surveys, this means that the impersonal screen interface could lead people to simply select the same answer for every question, something they could not do with a human interviewer. Chang and Krosnick (2010) utilize a true random assignment to either self-administered computer surveys or interviewer-enabled RDD phone interviews. They find that online participants provide fewer straightlining answers than RDD participants. Contradicting these findings, Heerwegh and Loosveldt (2008) uncover more 'straightlining' in web-based surveys than face-to-face. Similarly, Fricker et al. (2005) find evidence for more 'straigthlining' in web-based experiments than through the phone. Speeding concerns responses that were given 'too fast', given the number and depth of survey questions. Most research shows that online surveys display more speeding than other modes (Chang and Krosnick, 2010; Miller, 2000; Heerwegh and Loosveldt, 2008; Greszki et al., 2015). It is important to note, however, that speeding might not necessarily be bad for response quality. It could simply be that a visual mode of seeing the questions enables people to answer with the same quality, just more quickly than merely hearing the questions. Speeding research has yet to develop a method to distinguish these aspects.

Interviewer-induced measurement error occurs when the interviewer influences how re-

spondents answer the questions. The most common interviewer-induced measurement error is social desirability bias. The social desirability hypothesis proposes that in the presence of an interviewer, some participants may be reluctant to admit embarrassing attributes about themselves or may be motivated to exaggerate the extent to which they possess admirable attributes (Chang and Krosnick, 2010; Rogers et al., 2005). This body of research is consistent with the notion that self-administration by computer elicits more honesty (Kreuter et al., 2008). In other words, people are more honest when they answer sensitive questions on a screen than when they talk to another human being directly (be that face-to-face or on the phone). Other potential sources for interviewer-induced measurement error include gender, race, and age (Huddy et al., 1997; Rooney et al., 2005; Tourangeau and Yan, 2007; Warnecke et al., 1997; Yan and Tourangeau, 2008).

2.2.4 Nonresponse Error

Nonresponse error applies to the nonobservational gap between the sample and the respondent pool (Groves et al., 2009). Nonresponse error can occur on two levels, namely the unit and the item. Unit nonresponse error arises when a respondent who was supposed the answer the survey does not do so. Item nonresponse error happens when a respondent does not answer all questions or provides 'Don't Know' as an answer to one or several questions. It is very common for response rates to be very low, often in the area of 10 percent (Groves et al., 2009). This need not necessarily be worrisome, as long as these 10 percent look like the other 90 percent in terms of their covariates. It is only when differences between respondents and nonrespondents are systematic that nonresponse error comes into play. A high nonresponse rate does not equal the existence of bias, the same way a low nonresponse rate does not protect from bias. However, it is generally thought that a low nonresponse rate is preferred because it reduces the risk of nonresponse bias while not offering complete protection (Weisberg, 2005).

An abundance of studies on nonresponse in connection with survey mode exist. Atkeson

et al. (2014) find significant differences of nonresponse between telephone and online surveys, yet also show evidence that both modes represent the underlying population well. Somewhat similarly, Nagelhout et al. (2010) uncover differences between a web and telephone sample but maintain that these differences were small and not consistently favorable to either survey mode. On the contrary, Couper (2001) finds significant differences in nonresponse rates across modes. He concludes that nonresponse is a crucial concern for online surveys, especially when compared to the alternative method of mail. On the other hand, Chang and Krosnick (2010) identify lower nonresponse rates in computer surveys when compared to those administered by phone. They thus tentatively suggest a potential advantage of the online mode over telephone administration. Yet contrary again, West et al. (2013) emphasize interviewers' ability to influence hesitant participants on the phone and face-to-face, which is not possible online. Dillman and Christian (2005) in turn point out that nonresponse in online surveys can be overcome by software setup that obliges participants to answer all questions. These severely differing findings indicate that is far from clear whether there is a systematic difference in the levels of nonresponse and whether such a difference results in differing levels of data quality (Groves and Lyberg, 2010; McNabb, 2013; Millar and Dillman, 2012; AAPOR, 2013).

2.2.5 Post-Survey Error

Once data collection is completed, researchers often apply various statistical methods to adjust and improve the obtained results. These undertakings are best described as post-survey "efforts to improve the sample estimate in the face of coverage, sampling, and nonresponse errors" (Groves et al., 2009, p. 59). Any error that arises in this stage, when the survey data is processed and analyzed, is called post-survey error. Post-Survey error concerns everything that happens to survey data after collection is finished and includes data cleaning, recoding, weighting, and modeling, among many others (Weisberg, 2005). Whilst applicable to all survey modes, weighting and modeling to create representative survey results is heavily

used and widely spread particularly in online surveys (Atkeson et al., 2014). Any of these procedures can introduce error to the survey analysis process and thus bias the estimate (McNabb, 2013).

2.3 Entropy

2.3.1 Entropy for Discrete Survey Responses

Research on mode effects within the TSE thus does not reveal much about systematic, distributional differences that might be caused by mode choice. In particular, "despite the widespread use of online panels, there is still a great deal that is not known with confidence" (AAPOR, 2010, p. 54). In an attempt to address this issue, Homola et al. (2016) develop an entropy measurement (Shannon, 1948) for discrete survey measures. They argue that traditional statistical measurements, such as the variance ($\text{Var}(X) = \frac{1}{n-1}\sum(X_i - \overline{X})^2$) and the median average deviation ($\text{MAD}(X)$ =median($|X_i - \text{median}(X)|$), appear unsuited to the task of assessing this aspect as they assume interval measured data. In particular, $\overline{X}$ and median(X) are said to not represent correct measures of centrality (Homola et al., 2016; Wang, 2008; Hu et al., 2010). Entropy, on the other hand, is a measure of variability in discrete survey responses (Shannon, 1948).

Entropy roughly translates to 'information' and is most often used in information theory to carry levels of information in a message. It represents a measure of the average amount of information that is required to describe the distribution of a variable of interest (Cover and Thomas, 1991). Mathematically, Homola et al.'s entropy measure for a given response is formed by empirically counting the observations in each survey mode and subsequently normalizing them into probabilities. This measure, H, is calculated by ?

$$H(X) = \sum_{i=1}^{n} p(x_i)\log_2(\frac{1}{p(x_i)})$$

where $p(x_i) = Pr(X = x_i)$ is the probability of the i^{th} outcome of X. The entropy of the

variable X is thus the product of the probability of outcome x_i and the log of the inverse of the probability of x_i, summed over all possible outcomes x_i of X.

Based on Shannon (1948), Homola et al. (2016) argue that entropy is the only function that satisfies three critical properties: (1) H is continuous in the discrete measure $p_1, p_2, ..., p_n$, (2) H is at its maximum and is monotonically increasing with n if the p_i are uniformly distributed, and (3) The first H equals the weighted sum of consecutive H values, $H(p_1, p_2, p_3) = H(p_1, 1 - p_1) + (1 - p_1)H(p_2, p_3)$. Additionally, entropy does not make assumptions about the probability distribution of the variability of uncertainty.

Applying their measure to the 2012 ANES, which used identical questions that were asked face-to-face and online, Homola et al. (2016) find that entropy catches differences in mode effects that traditional statistical measurements that assume continuous data, e.g. the standard deviation, miss: Mode effects do not show in measures of centrality, but instead are shown by greater variability. They show that entropy reveals differences in how respondents react to identical questions presented in two differing modes. Their measure does not, however, account for ordinal variables, which are crucial in political science research.

2.3.2 Entropy for Ordinal Survey Responses

Why do we need an entropy measurement that accounts for ordinal variables? Homola et al. (2016) demonstrate that entropy reveals a measurement error in the form of a critical distributional difference in aggregate survey responses that traditional statistical measurements miss. Traditional statistical measurements also suppose ordinal data with roughly equally spaced distances between levels. Popular political science surveys, such as the ANES and the majority of survey experiments, however, contain ordinal data which are not equally spaced. Johns (2005), for example, points out that response scales that measure support or opposition often show significant differences in terms of spacing. On a 5-point Likert scale, the neighboring points 2 ("Somewhat oppose") and 3 ("Neither favor or oppose") are much closer together than the neighboring points 1 ("Strongly oppose") and 2. Ordinality

is a crucial feature of these variables. Treating ordinal variables as numerical results in the loss of this information of two crucial variables that predict political behavior. The lack of ordinality in an entropy measurement thus excludes two of the most important predictors of political behavior from survey mode effects analysis.

To develop a measure of ordinal entropy, I combine the Wilcoxon Signed Rank Test (WSRT) and Shannon's entropy measure to obtain a comparative measure of two ordinal vectors. The WSRT considers ordinal data as 'ranks'. By definition, ranks are not affected by outliers, since no outlying value is more than one unit away from the next value. The WSRT removes information about the distributional shape of the data. It is a statistical hypothesis test to compare two related samples to assess whether their population mean ranks differ, i.e. a paired difference test. It is used to determine whether two dependent samples were selected from populations that show the same distribution.

The WSRT follows several concrete steps. Let the first sample be denoted by units $x_{11}, x_{12}, ..., x_{1n}$ and the second sample by units $x_{21}, x_{22}, ..., x_{2n}$. We pair the units across samples, $(x_{11}, x_{21}), (x_{12}, x_{22}), ..., (x_{1n}, x_{2n})$, and define the paired differences, $d_1 = x_{11} - x_{21}, d_2 = x_{12} - x_{22}, ..., d_n = x_{1n} - x_{2n}$. We obtain the absolute value and sign of the paired differences and subsequently rank them, with values equaling zero being discarded. Let us call these values $R_1, R_2, ..., R_{n'}$, where the new $n' \leq n$ because of the discards that were carried out. We then calculate:

$$T^+ = \left| \sum_{i=1}^{n'} \text{sign}(d_i) \times R_i \right| = \left| \sum_{i=1}^{n'} \text{signed rank of } d_i \right|$$

The null hypothesis of the WSRT is H_0: median$(\delta) = 0$, with the alternative hypothesis being H_A: median$(\delta) \neq 0$. The new ordinal entropy measure I propose is simply the combination of the WSRT and Homola et al.'s entropy measure:

$$T^+ = \left| \sum_{i=1}^{n'} \text{signed rank of } d_i \times \log_2\left(\frac{1}{p(d_i)}\right) \right|$$

In the original WSRT, the signed rank of d_i are defined as the signed ranks of all paired absolute differences d_i. In our modified case, the signed rank of d_i are simply the signed ranks of the paired absolute differences between the two ordinal vectors of the two mode samples. HOW ARE THE 2 SAMPLES PAIRED? DOES THIS SATISFY HOMOLA'S 3 CONDITIONS? DO WE CARE?

3 Data

I apply my new method of ordinal entropy to external and original data in the online and face-to-face modes. For both purposes, I intend to conduct parallel studies with identical questions in the same period of time. I calculate ordinal entropy measurements for each survey mode. This reveals systematic, distributional differences that might be caused by mode choice. I then contrast these measures with the most common measure of the TSE for measures with validated baselines and the total survey measurement variation approach for questions on opinions and and attitudes. The most common measure of the TSE is the mean squared error, i.e. the squared average deviation of a survey estimate from the true value of the parameter being estimated. A large mean squared error indicates "that one or more sources of error are adversely affecting the accuracy of the estimate" (Biemer, 2010b, p. 826). For measures without validated baselines, i.e. questions regarding opinions and attitudes that lack previous survey exposure, I estimate the extent to which the surveys produce diverging measures of each concept in the respective survey modes (Smith, 2011).

3.1 External Data

External data sets with identical questionnaires face-to-face and online modes are provided by the 2012 ANES, as used by Homola et al. (2016), the 2016 ANES, and several Pew surveys. The Pew Research Center in particular has excelled in gradually shifting its modes. This provides an ideal testing ground to prove the validity of the ordinal entropy measure. I demonstrate that ordinal entropy matters when switching survey modes, as the new measure reveals discrepancies that cannot be detected with traditional statistical measurements (see

section 2.3.2).

3.2 Original Data

In addition to external data, I use original data to demonstrate ordinal entropy discrepancies and why they are significant. For this purpose, I field a survey experiment on the power of moral arguments online on MTurk and face-to-face for undergraduate students at American University. The details of this experiment can be found in paper III below.

Paper III: Moral Arguments as a Source of Frame Strength

1 Introduction

Barack Obama presented the first outline of the Affordable Care Act in the summer of 2009. The content of the reform was put online for everyone to see, but since the administration was still working on details, it refrained from actively communicating it. Published press releases simply stated that the ACA would expand coverage and lower health care costs for everyone. This hesitancy turned out to be a big mistake. At the end of July, support for the ACA hovered around 43 percent. Then Sarah Palin, John McCain's choice for running mate in 2008, posted the following statement on Facebook on August 7th: "The America I know and love is not one in which my parents or my baby with Down Syndrome will have to stand in front of Obama's 'death panel' so his bureaucrats can decide, based on a subjective judgment of their 'level of productivity in society'," (Palin, 2009). Palin implied that federal government workers would be able to refuse treatment to any patients and thus 'decide their fate'. Over the next two weeks, support for the ACA dropped to 35 percent while opposition rose to 52 percent. Republican lawmakers jumped at the opportunity and repeated the claim of 'death panels' whenever possible. The reform never recovered from this drop. In December 2009, four months after the statement, support and opposition were virtually identical to August. While the public was still uncertain about the exact contents of the law, Palin had asserted that it would include a Big Brother type panel that decided whether people would live or die. This drowned out any efforts by the Obama administration to show the law as a cost-reducing reform. Palin's frame of the ACA, in other words, drastically influenced public opinion of the reform.

Framing is the practice of presenting an issue to affect the way people see it (Aaroe, 2011; Druckman, 2001a; Gross, 2008). We learn about healthcare reform through articles, reports, speeches, commercials and social media. This mediated communication possesses tremendous potential influence on our perception of political issues (Iyengar, 1996; Kam

and Simas, 2010; Tversky and Kahneman, 1981). Framing research has established that a variety of frames substantively influence how people view and think about issues, such as the ACA (Price et al., 1997; Andsager, 2000; Callaghan and Schnell, 2005; Entman, 1993, 2004; Gamson and Modigliani, 1989; Lahav and Courtemanche, 2012; Pan and Kosicki, 1993; Slothuus and Vreese, 2010; Sniderman and Theriault, 2004; Vreese, 2004). But we do not know why these frames elicit these effects. A major challenge for framing research thus "concerns the identification of factors that make a frame strong" (Chong and Druckman, 2007, p. 116).

I conduct a series of tests to examine whether moral arguments form a part of what makes a frame strong. These tests include a meta-analysis, three online polls, and a survey experiment that is fielded online and face-to-face.

2 Theory

2.1 Framing

Despite the mass of experimental framing research, we still have little insight into what makes a frame strong. The larger persuasion literature "is not as illuminating as one might suppose... It is not yet known what it is about the 'strong arguments'... that makes them persuasive" (O'Keefe, 2002). One research direction is that frames are stronger overall when they cohere with an individual's personal value system. Feinberg and Willer (2012) frame environmental issues as a matter of 'purity', a theme that supposedly correlates with conservative ideology, and find this approach leads to increased conservative support of environmental policies. (Arceneaux, 2012, p. 280) on the other hand finds that "individuals are more likely to be persuaded by political arguments that evoke cognitive biases". Particularly, he asserts that messages which highlight out-group threats resonate to a greater extent than other, more coherent, arguments. In a study investigating the use of scientific data, Druckman and Bolsen (2011) report that adding factual information to messages about carbon nanotubes does nothing to enhance their strength. Providing more scientific evidence

seems to have the opposite effect, making the messages weaker. Overall, "it seems as if frame strength increases with frames that highlight specific emotions, invoke threat against one's own group interests, contain some incivility, include multiple, frequently appearing, arguments, and/or have been used in the past" (Druckman et al., 2018, p. 22). I attempt to provide an avenue of clarification by testing whether moral arguments are part of what makes frames strong.

2.2 Moralization

Moralization literature conceptually defines moral arguments as (1) near-universal standards of truth, (2) almost objective facts about the world, and (3) independent of institutional authority (Skitka, 2010). Moral arguments are said to engage a distinctive mode of processing that invokes a whole range of emotions and to be distinct from other forms of arguments (Ryan, 2014b). Scholars find that moral arguments are ubiquitous in political issues because they are essential to how people perceive and make sense of the world around them (Frank, 2005; Mooney, 2001; Tatalovich et al., 1994). Ryan (2014b) finds evidence that some people perceive distinctly economic issues such as labor relations laws or social security reform in moral ways. Other studies similarly assert that the strength of attitudes meaningfully differs when they are held with moral conviction (Baron and Spranca, 1997; Bennis et al., 2010; Ditto et al., 2009; Tetlock, 2003). It is also asserted that moral conviction represents an important force that guides citizen behavior and development of public opinion (Converse, 1964; Skitka et al., 2005; Skitka and Wisneski, 2011; Smith, 2002; Tatalovich and Daynes, 2011; Zaller, 1992). It is widely argued that people rely to a disproportionate extent on moral arguments to form their opinions and apply this moralization to political issues (Ryan, 2014a,b; Smith, 2002). Moral arguments can achieve a much higher emotional connection with people because they invoke people's values and feelings (Skitka et al., 2005; Skitka and Wisneski, 2011; Haidt, 2003; Tatalovich and Daynes, 2011). These conceptual definitions are all encompassed in Moral Foundations Theory (MFT), developed by Haidt (2012) and

presented in Table 2 below.

Table 2: Foundations of Moral Arguments

Positive		Negative	
Care	Cherishing, protecting others	*Harm*	Hurting others
Fairness	Rendering justice by shared rules	*Cheating*	Flouting justice, shared rules
Loyalty	Standing with your group	*Betrayal*	Opposing your group
Respect	Submitting to tradition, authority	*Subversion*	Resisting tradition, authority
Sanctity	Repulsion at disgust	*Degradation*	Enjoyment of disgust

Based on Haidt (2012). Positive and negative foundations are conceptual opposites.

These aspects of moralization theory have not been applied to frame strength in experimental research. An abundance of framing research has shown that frames elicit significant changes in issue positioning (Chong and Druckman, 2013, 2010; Druckman et al., 2013; Druckman, 2001b; Druckman et al., 2012; Nelson et al., 1997; Slothuus, 2008). Brewer and Gross (2005), for instance, find significant effects for the frames 'School vouchers create an unfair advantage' and 'School vouchers provide help for those who need it'. Druckman et al. (2012) provide similar evidence for 'The Affordable Care Act gives more people equal access to health insurance' and 'The ACA increases government costs', while Druckman et al. (2013) do so for 'Oil drilling provides economic benefits' and 'Oil drilling endangers marine life'. While we know these frames elicit changes, we do not know why they do so. We do not know why these frames 'work'. I propose a way to understand why some of these frames work, which is based on moralization theory: The presence of moral arguments.

Applying Moral Foundation Theory, both frames in Brewer and Gross (2005) could be categorized as containing moral arguments, even though the authors do not explicitly do so. 'School vouchers create an unfair advantage' can be argued to contain the negative moral foundation of *Cheating*, while 'School vouchers provide help for those who need it' contains the positive moral foundations of *Care* and *Fairness*. 'The ACA gives more people equal access to health insurance' (Druckman et al., 2012) also could be said to contain the positive moral foundations of *Care* and *Fairness*, while 'Oil drilling endangers marine life' (Druckman

et al., 2013) contains the negative moral foundation of *Harm*.

It is important to note that 'The ACA increases government costs' (Druckman et al., 2012) and 'Oil drilling provides economic benefits' (Druckman et al., 2013) do not directly appeal to morality, yet research has shown them to be strong. Of course, some people might see increasing costs immediately as bad and thus morally detrimental for the future of the country, but these frames do not make such an appeal directly, on the surface. This distinguishes them from moral frames.

This might lead one to assert that moral arguments do not form a part of frame strength – after all, if a frame does not contain a direct moral argument but is proven to be strong nonetheless, surely then frame strength does not depend on the presence of moral arguments. This hypothetical argument is flawed, however. For one, the search for the source of frame strength is not the search for a universal 'holy grail' argument whose presence is a precondition for frame strength. There probably are many aspects that can make a frame strong, with moral arguments potentially being one of them, not the only one. Second, the studies with these two amoral and two moral arguments do not distinguish between the directions in which the moral and amoral frames act.

In the case of Druckman et al. (2012), 'The Affordable Care Act gives more people equal access to health insurance' contains a positive, i.e. support-inducing, moral argument, while 'The ACA increases government costs' contains a negative, i.e. opposition-inducing, amoral argument. They act in opposite directions. A comparison of these frames alone does not yield sufficient results as we would not be able to identify the exact cause of the framing effect. Is 'The Affordable Care Act gives more people equal access to health insurance' strong because it supports the issue or because it contains a moral argument? Similarly, is 'The ACA increases government costs' strong because it opposes the issue or because it contains a amoral argument? This set-up cannot answer these questions.

To establish whether moral arguments are part of what makes a frame strong, we need a design that assesses the strength of frames with moral and amoral arguments whilst ac-

counting for both signs, opposing and supporting, in both sets of frames. I provide such a design.

Overall, experimental framing research has shown that many frames have significant framing effects. It is still unclear, however, what makes these, or indeed any, frame strong (Druckman et al., 2018). Moralization theory claims that moral arguments possess enormous power shape human behavior and influence public opinion (Haidt, 2003). I combine these two sets of literature and analyze whether moral arguments form a part of what makes frames strong. I use a variety of data sources to investigate the following hypothesis:

H. Moral arguments form a part of what makes political frames strong.

3 Data

First, I conduct a meta-analysis of all experimental framing surveys up to the present day. The aim of this step is to obtain statistically valid estimates of all the frames that have been shown to elicit significant framing effects. Second, I field an online poll that asks participant to assess how moral these researched frames are. Third, I conduct a second online poll to obtain qualitative insights into what people consider to be moral arguments. Fourth, I design and test moral and amoral complementary frames to the ones used in previous research in a third online poll, based on insights from the meta-analysis and the two previous polls. Finally, I conduct the main online survey experiment.

The insights gained from the second online poll are vital to the designing of the frames. Possessing knowledge of the way people define moral and pragmatic arguments is crucial to the development of frames that accurately reflect these structures. To add a second layer of insurance to protect against miss-designed frames, I also pre-test the designed frames with participants who are not part of the main survey. These participants are exposed to the designed frames and asked whether they reflect the core ideas behind moral and pragmatic arguments. This pre-test structure builds on work by Slothuus and Vreese (2010) and Chong and Druckman (2007), following the mass communication and persuasion literature (O'Keefe,

2002). The pre-test is carried out on Amazon's online platform MTurk. Together with the post-test included in the survey, the assessment of moral arguments and a pre-test represent a thorough, robust, spread-out safety net to ensure that the designed frames connect with participants, which in turn ensures meaningful survey results.

3.1 Meta-Analysis

A meta-analysis "synthesizes the findings of many different scientific studies of the same phenomenon" through statistical analysis (Groves et al., 2009, p. 190). The basic tenet behind meta-analyses holds that there is a common truth behind all conceptually similar scientific studies, but this truth has been measured with a certain error within individual studies (Schmidt and Hunter, 2015; Flores, 2015; Greenland and O'Rourke, 2008). A key benefit of this approach is the aggregation of information leading to a higher statistical power and more robust point estimate than is possible from the measure derived from any individual study (Walker et al., 2008). In performing a meta-analysis, an investigator must make choices which can affect the results, including deciding how to search for studies, selecting studies based on a set of objective criteria, analyzing the data, and accounting for or choosing not to account for publication bias (Brockwell and Gordon, 2001; Paterson, 2001; Kuhberger, 1998; Boulianne, 2009).

The first step in a meta-analysis is a comprehensive inventory of the body of research. I use the large number of empirical framing studies I analyzed over the years and supplement this by searching public registries, such as *E*[illegible]. To account for publication bias, I also include pertinent unpublished papers.

The second step is to develop criteria for inclusion. I only include research with a focus on experimental framing. This means that studies on related, but different, concepts such as priming or agenda-setting and studies of a descriptive nature are excluded from the data.

The third step involves the calculation of effect sizes.There may be areas of heterogeneity between the respective studies, such as differing selection and treatment of subjects, or

differences in the conception of outcomes or policies (Roever, 2017). This can be overcome by using a Bayesian random effects model. In the Bayesian paradigm, the unknown parameters involved in the model are considered random variables (Gelman et al., 2013). After eliciting a priori distributions, these priors are combined with the empirical evidence (likelihood) to obtain a joint posterior distribution, given the data we can observe. Markov chain Monte Carlo (MCMC) methods can be used to obtain a sample from the joint posterior distribution of the parameters. We use the summaries of the marginal posterior distribution of the parameters of interest (i.e. in our case group means) to leverage inference (all from Gill, 2014). The Bayesian approach accounts for the uncertainty around the heterogeneity variance. Frequentist approaches (e.g. the t statistic), on the other hand, treat the estimate of the variance as a fixed quantity, which results in variability underestimation (Lan et al., 1999).

3.2 Online Poll #1: The Moral Content of Frames Used in Previous Studies

Once the meta-analysis is completed, I will have a statistically sound assessment of the framing effects found in previous research. I will also have a list of all significant frames used in previous research. I use these frames in a simple online poll, which asks participants to rate how moral each argument in each frame is on a scale of 1 (Not moral at all) to 5 (Very moral). This provides me with a list of how moral or amoral the strong frames from previous research. The poll is fielded on MTurk.

Because the design in previous experimental framing studies does not account for direction (see section 2.2), this list is a mixture of positive moral frames, negative moral frames, positive amoral frames, and negative amoral frames. Crucially, there will not be 'perfect pairings' that account for direction of support and moral content. To fill these missing 'slots', I design moral and amoral frames that complement the existing moral and amoral frames. Depending on what the missing 'slot' is, these frames are either positive moral, negative moral, positive amoral, or negative amoral. In order to do so, however, I first conduct another online poll, insights of which inform the design of said frames.

3.3 Online Poll #2: What People Consider Moral Arguments

The aim of this online poll is to obtain qualitative insights into what people consider to be moral arguments. Survey questions are best designed if researchers have a firm grasp of the underlying realities that participants have to report (Carpini and Keeter, 1993; Stanley, 2016; Conover et al., 1991). It is vital to understand how target populations understand and construct key concepts in the eventual survey questionnaire (Groves et al., 2009). If I want to accurately design complementary moral and amoral frames as the next step, I first need to know how people construct this concept. A separate online poll provides these insights to inform the final survey research project.

3.4 Online Poll #3: The Moral Content of Designed Complementary Frames

I use the insights from the second online poll to design several frames for each of the missing 'slots' in previous experiments. I develop several frames for each missing 'slot'. Once I have designed these frames, I will conduct another simple online poll which asks participants to rate how moral each argument in each designed frame is on a scale of 1 (Not moral at all) to 5 (Very moral). Similar to the online poll in section 3.2, this assesses the moral content of the frames I designed. This is to make sure that the designed frames connect with people in terms of moral content in the intended way. Like the previous poll, this poll is also fielded on MTurk.

3.5 Survey Experiment

Finally, I design a questionnaire for a survey experiment that combines all the insights from the previous tests and thoroughly examines whether moral arguments form a part of what makes a frame strong. The experiment is fielded twice, once online for a random sample of U.S. adults on MTurk, and once for a random sample of undergraduate students at American University. It features frames with moral and amoral arguments that have been shown to elicit effects (revealed by the meta-analysis and tested for moral content in the first online

poll) and frames with moral and amoral arguments that fill missing 'slots' (designed based on insights from the second online poll and tested for moral content in the third online poll). The survey also collects demographic information and includes post-treatment evaluations in terms of rating the arguments by their moral content. The political issues the frames apply are determined by the issues used in previous framing experiments. Both methodological contributions from paper I and paper II of this dissertation are applied in the online and face-to-face fielding of the survey.

3.6 Failsafe in Case of Null Results

By splitting the analysis into several steps from different data sources, I tried to minimize the potential impact of potential null results in the final online survey experiment. To my knowledge, a meta-analysis (section 3.1) of experimental framing studies has not been conducted. The assessment of the moral content of effective frames in previous experiments (section 3.2) also represents a significant undertaking in its own right, as it provides a type of test that has not been done in framing research on the subject of moral content. The insights from the second online poll (section 3.3) could also be incorporated by informing the discussion of moral content. Even if the eventual survey experiment reveals null results, then, these steps would still provide enough content for a publishable article.