
16. Experiments in political science

Damien Bol

Like other social sciences, political science has embraced an experimental turn in recent decades (Angrist and Pischke, 2010). Today, the leading journals within the discipline predominantly feature studies that use experimental designs to examine the effects of treatment X on outcome Y (Bol, 2019; Druckman and Green, 2021; Torreblanca et al., 2026). While descriptive papers are still published, they are now explicitly labeled as such, with authors being careful in how they interpret their findings. As a frequent reviewer, I can tell that political scientists, akin to their peers in other disciplines, cannot get away with suggesting that correlation implies causation in their research papers and hope to get published.

In the subsequent pages, I will explore the experimental method that has become particularly prevalent in political science: survey experiments (Thomas, 2024; Torreblanca et al., 2026). This method involves conducting an experiment within a survey, nowadays mostly online, with a large sample typically representative of a given country's population. It has garnered considerable interest due to its compatibility with the discipline's focus on context-dependent phenomena. Unlike economics or psychology, political science's objects of study tend to be tied to certain environments. Voting behavior, for instance, takes a different form depending on the electoral system used in the country and the parties that compete for elections. Even at the individual level, engagement with elections varies considerably depending on personal characteristics like social status. For these reasons, political scientists are often skeptical that patterns observed in small convenience samples of subjects like university students can be generalized to a large and more diverse population (Kam et al., 2007). That is why they usually rely on large samples representative of the population of a country, or even ideally from several countries, to ensure the external validity of the results, their relevance, and meaningfulness (Bol, 2024).

In this chapter, I will thus discuss survey experiments and provide some practical advice for researchers interested in employing the method. In the first section, I will outline this category of methods, spotlighting recent progress in the field. I will then discuss a series of “dos and don'ts” for interested researchers, focusing on the use of monetary incentives, the incorporation of manipulation and attention checks, the determination of the sample size, and considerations about the location of the experiment within the broader survey questionnaire. Towards the end of the chapter, I will briefly touch on other experimental approaches in political science, such as field and lab experiments, offering references for those keen on exploring these methodologies. Finally, I will address some limitations of survey experiments and suggest strategies to overcome them.

WHAT IS A SURVEY EXPERIMENT?

Political scientists, like sociologists, have long engaged in surveying populations. Since the early days of the discipline, they have asked political questions to large samples of ordinary citizens, often representative of a country's adult population or that of a smaller geographical area. They inquire about respondents' opinions on various policies, their feelings toward the current government, or whom they voted for in the last election (e.g., Lazarsfeld et al., 1948 represents one of the earliest studies of this kind). The method has become so widespread that nearly every country around the world is surveyed multiple times per year by political scientists. In this context, it was only natural for them to start including experiments in these surveys when the experimental revolution swept through the discipline.

At its most basic level, survey experiments involve dividing survey respondents randomly into a control group on the one hand and one or more treatment groups on the other.¹ Researchers present some piece of information to those in the treatment groups during the survey, whereas they do not show anything to those in the control group. These treatments often consist of a brief text, an image, or even a video. Sometimes, respondents from the control group are exposed to some unrelated information bit that serves as a placebo when the researcher expects that mere exposure to any type of information will affect answers to subsequent survey questions. The researchers then compare the responses to a subsequent survey question, which constitutes the outcome, from those in the control group with those in the treatment group or groups. Because of randomization, any differences in outcome can be attributed solely to the treatment. A simple t-test or bivariate OLS regression is then used to calculate the average treatment effect.²

These studies, often referred to as vignette experiments, are prevalent in political science. The ease of conducting such experiments has been enhanced by the advent of web-based surveys, where respondents answer questions in privacy on their computers, phones, or tablets. Nowadays, many polling companies like Nielsen, YouGov, or Ipsos offer very attractive prices for researchers who want to conduct a survey of five or ten minutes with large representative samples of citizens from a wide range of countries that are made representative of the population of these countries using socio-demographic quotas. In the conclusion, I discuss issues regarding the representativeness of these samples.

An example of vignette experiments includes the research by Samuels and Zucco (2014), where they queried a group of Brazilian respondents about their stance on various political issues, such as whether the minimum wage should be lowered (outcome *Y*). Prior to this, the authors randomly presented the main parties' positions on these issues to a subgroup of respondents (treatment *X*). By comparing the responses of those who received this additional information with those who did not, they were able to estimate the impact of awareness of the main parties' stances on issues, or treatment, on the likelihood of respondents aligning their views with one position over another, or outcome. They found that respondents were more inclined to align with their preferred party's position by about 15 percentage points when aware of it. From these findings, they deduced that partisanship plays a significant role, even in relatively new democracies like Brazil, as it influences individuals to shift their political views.

Another increasingly common type of survey experiment in political science is the conjoint study (Hainmueller et al., 2014). Originating from marketing sciences, these studies ask

respondents to choose between two hypothetical options, such as two public policies or two candidates in an election (outcome *Y*). The rationale is that making a choice between two options is relatively natural for many individuals and straightforward, cognitively speaking. The treatments are the set of attributes that the researcher uses to define these two options (treatment *X*); for a hypothetical candidate, this could include their age, gender, occupation, and party affiliation, for example. The key is that these attributes vary randomly across the options, allowing the researcher to estimate the causal effect of each attribute on the likelihood of a respondent choosing one option over the other. Practically, this involves restructuring the dataset so that the unit of observation is the options presented to the respondents. The researcher can then regress the choice on the attributes of these options to estimate the average treatment effect.

An example of conjoint analysis is the study by Teele et al. (2018), in which the authors asked a representative sample of American citizens to choose between two hypothetical candidates in a fictional election. They varied attributes such as political experience, gender, marital status, and number of children, among others. Their analysis revealed that Americans do not exhibit prejudice against female candidates; in fact, they are more likely to favor women over men for the same election. The average treatment effect associated with being a female instead of a male candidate is indeed four percentage points and statistically significant, indicating that respondents are more likely to opt for female than male candidates, independent of all other attributes. The authors concluded that the underrepresentation of women in US politics is due to factors other than voter bias.

Some other experimental designs do not directly aim at estimating an average treatment effect but seek to enhance the accuracy of responses to sensitive survey questions. An example is list experiments (Blair and Imai, 2012), which are particularly useful for sensitive questions where respondents might be reluctant to reveal their true beliefs. In these experiments, both control and treatment groups are presented with a list of items, with the treatment group receiving an additional sensitive item, such as a statement about restricting the movements of Black people to predominantly white neighborhoods. Given the potential discomfort in agreeing with such statements openly, even in an anonymous survey, respondents might conceal their true opinions to adhere to perceived social norms.

The methodology behind list experiments involves not asking respondents to express their agreement with each item individually but to report the total number of items with which they agree. This approach potentially increases respondents' comfort in reporting their true feelings, as researchers cannot determine the specific items agreed upon. However, because of the randomization, any difference in the total number of items agreed upon by the groups can be attributed to the sensitive item. Researchers can then use this difference to estimate the proportion of respondents that harbor the specific attitude in question. For instance, Kuklinski et al. (1997) used a list experiment to estimate that about 40 percent of white US citizens harbored negative views toward affirmative action policies, with an even higher percentage in the southern states.³

Similarly, some studies leverage randomization to improve question wording in surveys, addressing issues like social desirability bias. For instance, respondents might inaccurately claim to have voted when they have not (Karp and Brockington, 2005). Morin-Chassé et al. (2017) found that incorporating a face-saving response option, allowing respondents to indicate "they usually vote but did not this time," reduces overreporting of turnout by about eight percentage points. This was determined by comparing the responses from a control group,

which answered a traditional yes/no voting question, to those from a treatment group asked a question including a face-saving option. The main distinction between these experiments and list experiments is that the randomization here is solely to identify which question version most accurately reflects reality. After determining this, researchers do not need to include an experiment in their surveys.

Note that some researchers use randomization in surveys for purposes other than conducting an experiment. This happens, for example, when they suspect that question ordering might influence responses. To avoid this risk, they randomize the order in which questions are asked to respondents or the order in which response categories appear on the respondent's screen. They do not need to control for this ordering unless this is part of a hypothesis they want to test. Randomization ensures that it cannot affect other results. Similarly, when researchers need to ask a large number of repetitive questions in a survey, they sometimes use randomization to minimize respondents' fatigue. For instance, when they want respondents to evaluate a long list of scenarios, like policies, they sometimes only ask each of them to evaluate a random subset of these scenarios. Here again, they do not need to worry about who evaluates which scenario as randomization ensures that each respondent has the same ex-ante probability to evaluate them. However, my suggestion for researchers is to avoid using the term "experiments" to describe these designs in which randomization is not central to the test of hypotheses in order to avoid any confusion (e.g., Blais et al., 2023).

Finally, note that although survey experiments have traditionally been conducted with samples of ordinary citizens, a new trend in political science is to conduct survey experiments with samples of politicians, usually locally elected officials or candidates, to have a sufficient number of observations (e.g., Esaiasson and Öhberg, 2020). This evolution fits the common approach in the discipline to study its objects in context (see section above). Indeed, several studies have found differences in results when they replicate the same survey experiment with comparable samples of politicians and citizens. For example, they found that politicians are ten percentage points more likely to succumb to the sunk cost fallacy than citizens (Sheffer et al., 2018). Importantly, these elite surveys are much more difficult to conduct, as they usually require the researchers to invest much time to convince politicians to participate in surveys in the first place. As far as I know, no polling company provides this service at the moment.

DOS AND DON'TS OF SURVEY EXPERIMENTS

In this section, I will provide some concrete advice to researchers interested in conducting survey experiments like those conducted in political science. I will cover the following topics: the use of monetary incentives, the incorporation of manipulation and attention checks, the determination of an appropriate sample size, and the location of the experimental part in the broader survey questionnaire.

Monetary Incentives

Many experimentalists in the social sciences, especially experimental economists, incorporate monetary incentives as part of their study design. At the conclusion of these studies, subjects receive some amount of money that depends, at least partly, on the decisions they made during the experiment. If they make "clever" choices, they earn more money than if they

do not. These incentives aim to ensure that subjects take the experiment seriously, thereby reducing noise and biases in the data. The concern is that without them, subjects might at best behave randomly or, worse, erratically in unpredictable and unrealistic ways. In this way, monetary rewards guarantee the meaningfulness of the subjects' behavior and the quality of the data (Camerer and Hogarth, 1999).

Political scientists generally avoid including monetary incentives in their survey experimental designs. Their apprehension, also shared by some experimental economists, is that in some contexts, these incentives might crowd out intrinsic motivations and thus alter behaviors (Villeval, 2007). This is particularly relevant in political science, where many of the phenomena studied are motivated by intrinsic factors, such as group identity. For example, when voters decide which party to support, they are often driven by moral or identity-based motivations, knowing that their individual vote is unlikely to significantly impact the final electoral outcome (Feddersen et al., 2009). In this sense, their choice is driven by an expressive drive to support the party that resonates with them, rather than an instrumental desire to influence the election results (Bol et al., 2018). For these reasons, unlike many online experiments, survey experiments do not usually include monetary incentives. There is a concern that doing so would distort the responses, which should be motivated by intrinsic factors, just as political behaviors and attitudes are in real life.

To be clear, most political scientists do employ monetary rewards for their survey experiments. However, they use them to motivate respondents to participate in the survey in the first place. All receive a fixed sum of money upon completing the questionnaire. Without this reward, response rates would be low, which would undermine the goal of having a sample of respondents that is representative of the population of the country studied, as only individuals with a deep interest in political surveys would participate. However, political scientists do not use monetary incentives within the survey; respondents do not earn more money depending on how they answer any of the survey questions.

Note, however, that the absence of monetary incentives in survey experiments poses a problem for data quality. Given that respondents can provide any answer in the survey and still receive the same monetary reward at the end, there is a risk that they do not engage with the survey much and respond in a random manner due to inattention. Worse, they might also systematically lie in their answers, for example, to sensitive questions for which there is a social desirability bias (see above and Chapter 15 by de Quidt, Vesterlund, and Wilson in the present volume). This is the cost of refusing to use monetary incentives and a reason why including attention and manipulation checks is crucial (see section below). In the future, it would be beneficial for political scientists to consider other types of incentives that better align with the nature of the subjects they study. Perhaps an interesting solution would be to include monetary incentives not directly given to the respondent, such as donations made on their behalf to a charity supporting a cause they endorse or ask respondents to perform a meaningful and time-consuming activity within the questionnaire, such as writing an email to a real politician urging them to address a certain policy issue (Adida et al., 2018). This approach would better match the identity-based motivations that lie at the core of the discipline but minimize the risk of inattentive respondents and meaningless answers.

Manipulation and Attention Checks

As political scientists generally do not use monetary incentives in their survey experiments, it becomes crucial to employ other strategies to ensure the quality of the data. Besides the issue of inattentive respondents and meaningless answers, a common problem with surveys conducted online, like most surveys nowadays, is the presence of “bots.” These programs fill out questionnaires on behalf of real respondents, allowing them to participate in numerous surveys at minimal cost and still earn the fixed reward. These bots behave quasi-randomly in surveys and thus introduce significant noise into the data (Westwood, 2025).

A first solution to guard against inattentive respondents (but not bots, unfortunately) is to resort to common sense and avoid long surveys of more than 10 minutes and questions of more than a few lines of text. This is especially important for the experimental part of the survey. For example, in a vignette experiment, a researcher cannot reasonably expect a respondent to read more than a handful of lines of text or watch a video of a few seconds. If it is too demanding, the respondent will most likely skip it, defeating the purpose of the experiment. Therefore, I generally recommend designing short surveys lasting between 5 and 10 minutes (or less) and imagining treatments that are almost effortless for the respondent to be exposed to. This can be a figure or a table presenting the information in a way that is easy to read. In the case of a conjoint study, that would consist of setting the number of attributes that define the two options to five or six. The rule here is not to overload the respondent’s capacity and generate undue fatigue. For example, in their survey experiment about Brazil, Samuels and Zucco (2014) constructed short vignettes in which the respondents saw on their computer screens where the two main parties in Brazil stand on a set of political issues. Each time, the information was communicated in a single sentence, easy for respondents to read.

Moreover, the two most common strategies to ensure the quality of the data are to include attention and manipulation checks (Kane, 2024). An attention check aims to identify and exclude bots and inattentive respondents who answer the survey randomly. It involves adding a question, usually at the beginning of the survey or at least before the experimental part, with a clear correct answer, easily identifiable by anyone who reads the question. For example, this could be a question asking the respondent to select a given number, say “3,” from a list containing various numbers like “1, 7, 9, 3, 4, and 2.” The rationale is that inattentive respondents and bots will not read the question and will select a random number, which has little chance of being the correct one. The researcher can then exclude those who give the wrong answer to this question and hopefully ensure that most remaining respondents are real and attentive.

A second strategy is to include a manipulation check, which measures whether the treatment exposure moves the beliefs and attitudes of respondents in the expected way. It involves adding a question, right after the experimental part, on a topic related to the content of the treatment. This question is almost like a comprehension test, for which the researcher expects clear differences in answers between the control and treatment groups. A key challenge with the manipulation check is to find a question that does not feel repetitive or redundant for the respondents who have just been exposed to the treatment.

For example, in a survey experiment about public support for citizens’ assemblies in Europe, Pilet et al. (2023) presented a vignette showing the real percentage of individuals who have a certain position on a given issue, like immigration, e.g., “55 percent of the population thinks that immigrants have an overall positive effect on the national economy,” to a random group of respondents. The control group did not receive this information. To ensure that respondents

read the content of this treatment, they included a manipulation check right after it, asking respondents to guess the share of the population that supports a related, yet slightly different statement to avoid repetitions. They asked: “In your opinion, what share of the population thinks that the country should not accept new migrants?” They find that most respondents in the treatment group propose a guess close to the benchmark of the vignette (e.g., 55 percent), unlike those in the control group, whose guesses are all over the place, suggesting that those who received the extra bit of information contained in the treatment updated their beliefs about their compatriots accordingly.

With manipulation checks, the researcher can thus identify whether the treatments “worked” as intended and potentially identify for whom they did and did not work. In every survey experiment, a substantial portion of respondents are not affected by the treatment, in the sense that they did not update their beliefs or attitudes after being exposed to it, due to inattentiveness or other reasons. These respondents are usually referred to as “non-compliers.” The researcher needs to accept this limitation, as reducing the sample in the treatment group to compliers would introduce a post-treatment bias in the estimates. Indeed, respondents who update their beliefs or attitudes after being exposed to the treatments are likely to be different from those who do not. Thus, removing this type of respondent from the treatment group’s sample would create an imbalance between the treatment and the control group. Indeed, the researcher would not be able to identify the non-compliers from the control group, which means that it would cease to be a good counterfactual for the reduced treatment group.⁴ For these reasons, survey experiments necessarily create “intention-to-treat” effect estimates, i.e., ITT in experimental jargon, rather than effective average treatment effects, which correspond to situations in which every respondent from the control and treatment group is a complier. In this sense, estimates from survey experiments are underestimations or lower bounds of the “real” treatment effects.

Besides ensuring that treatment exposure moves respondents’ beliefs or attitudes in the expected way, manipulation checks are useful for at least two other reasons. First, they constitute another form of attention checks, giving the researcher an indication of whether respondents were attentive or not, or at least sufficiently attentive to notice the treatment. Second, manipulation checks help identify treatments that are not effective because the extra bit of information they convey is already known to respondents before the experiment begins. For example, in their study about citizens’ assemblies, Pilet et al. (2023) replicated the vignette experiment about immigration for another political issue: social benefits. Here, they found little difference between the real share of the population supporting the issue and the guess made by respondents, even in the control group. This suggests that respondents already had some ideas about the actual number even before this experiment. They did not need to be exposed to the information contained in the treatment to formulate a correct guess. In this situation, the treatment cannot have an effect as it did not move respondents’ beliefs or attitudes. This leads to another important piece of advice for researchers interested in survey experiments: they should imagine a treatment that provides an extra bit of information that is relatively unknown to respondents before the start of the study. Otherwise, the treatment is not about moving their beliefs or attitudes but about reminding them of this information, or “priming” in experimental jargon, which is a completely different objective.

Determination of Sample Size

As mentioned above, the *raison d'être* of survey experiments is to encompass a large sample of respondents, usually representative of the population of a certain country or countries. To achieve representativeness, polling companies utilize census information about the actual socio-demographic composition of these countries. When researchers request that an online questionnaire be sent to a representative sample of the population, they instruct their panel, i.e., individuals who agreed to participate in studies for them, to fill it out. They ensure that the number of respondents in each socio-demographic stratum of the population, e.g., middle-aged white women with a university degree, is proportional to their representation in the actual population of the country.

The advantage of using large, representative samples of the population is twofold. First, it aligns with the overall approach of political science, a discipline that considers most phenomena it studies to be context-dependent. With a representative sample, researchers can be confident that the average treatment effect is not due to the over-representation of some types of respondents who are particularly sensitive to the treatment's exposure. In fact, it is increasingly common for studies based on survey experiments, where the sample is not perfectly representative of the population, for instance, because the response rate among contacted respondents is lower in some socio-demographic strata than in others, to recalculate the estimates by weighting the data based on census information (Egami and Hartman, 2023). This reverse engineering exercise involves giving more weight to the strata that are under-represented in the sample compared to their actual prevalence in the population.

The importance given to context in political science has another implication for the analysis of survey experiments. Beyond calculating the average treatment effects over the full sample of respondents, researchers generally also estimate the average treatment effects across sub-samples, often socio-demographic groups, e.g., among men and then women, or among young and then older individuals. They do so either by simply splitting their sample and reproducing the same OLS regression or by adding an interaction between the treatment variable and the variable defining the socio-demographic group. The rationale behind this strategy, often called "heterogeneous analysis," is to explore whether the average treatment effect is universal across all groups or driven by specific groups. Given political science's focus on context-dependency, this question is crucial to many studies. Researchers thus need to consider this objective when performing their power analysis, ensuring enough observations in each social group to detect a statistically significant effect. This requirement is especially relevant when subgroup analysis is central to the study, typically when it constitutes one of the hypotheses.

The second advantage of relying on a large sample of respondents is that the researcher can detect even small average treatment effects. This is crucial for survey experiments where respondents' answers are not incentivized by monetary rewards. As mentioned earlier, this tends to create noise in the data due to inattentive respondents, bots, and non-compliers. For this reason, survey experimentalists usually calculate, *ex ante*, the appropriate sample size needed to detect their effect, an operation called a "power analysis." From a likely average treatment effect based on findings from previous studies on similar topics, e.g., an increase of seven percent of the standard deviation of the outcome variable, they deduce the number of observations in each group required for this effect to be statistically significant at a level of $p < 0.05$ or 0.01 , usually with a power level of 80 percent (Arel-Bundock et al., 2024). Most data and statistical analysis software nowadays include a tool to perform such a calculation.

Such power analysis is usually central in pre-analysis plans. These plans are documents that researchers circulate online on public repositories like the Open Science Foundation (see <https://osf.io/3wct2/wiki/home>) before they even start collecting data. They are increasingly common in political science as many top journals now require submitted papers based on any kind of experiments to have one. A growing number of journals even allow researchers to submit a pre-analysis plan (or registered reports) in place of a paper. If the plan is accepted, the researchers then commit to conducting the experiment as announced, and the editors commit to publishing the full paper regardless of the results.

Pre-analysis plans usually contain a series of guidelines, typically the hypotheses that will be tested and how, i.e., the variables that will be used and the estimation technique (Nosek et al., 2018). The goal is to improve transparency and reliability in academic research by demonstrating that researchers are not engaging in “p-hacking,” which refers to the dishonest practice of performing numerous analyses until finding one that yields statistically significant results confirming their hypotheses. Similar to other types of experiments (see Chapter 18 in this volume by Dreber and Johannesson), with a pre-registration plan, survey experimentalists commit to reporting the analyses they envisioned before the start of the experiment, regardless of whether the results fit their argument.

A final element researchers should consider while determining the sample size is whether respondents will have the option to respond “Don’t know” or “Prefer not to answer” to questions capturing the outcome variables. These response categories are common in surveys in general. Aside from ethical concerns, including them poses a tradeoff for researchers. On one hand, by not allowing respondents the option not to answer, they increase their sample size. On the other hand, by providing this option, they risk introducing even more noise or bias into the data. Indeed, when respondents genuinely do not know how to respond, they might choose a random response category or systematically make their choice based on the first attribute only, without considering the others.

This tradeoff becomes evident in conjoint studies where the respondent chooses between two hypothetical candidates. In real elections, voters can usually decide to abstain and not choose any of the competing candidates. This is even more likely to happen in a survey experiment where the candidates are hypothetical, as the respondent does not have the time to develop a strong opinion about them. This means the researcher needs to choose between allowing the option not to answer, risking too few observations to detect any average treatment effect, or forcing a choice, risking introducing too much noise into the data. This tradeoff also explains why conjoint studies typically ask respondents to repeat the choice task five or even ten times. Each repetition generates new options due to random attributes, increasing the sample size. Evidence shows that this strategy is effective as it does not necessarily introduce more noise or biases into the data (Bansak et al., 2018).

Where to Place an Experiment in a Survey?

By definition, survey experiments are conducted within surveys. Sometimes, the experiment constitutes the survey’s main focus. Other times, it forms only a small part of a longer survey. This occurs when multiple researchers pool their resources to conduct a single survey, using the collected data to address various research questions on different topics. This approach is common due to the economies of scale it offers. However, it raises an important question regarding the order of questions and treatments.

A crucial principle for researchers to remember is that any survey questions following the exposure to treatments must be considered as outcome variables. Given their placement in the survey, the responses from the treatment group to these questions could be influenced by exposure to this treatment. Researchers can then determine the average treatment effect on responses to these questions, but cannot use these responses to weight the data or to explore potential heterogeneous effects. Doing so would introduce a post-treatment bias, meaning that the treatment group's responses to these questions, likely influenced by the treatment, would no longer be comparable to the control group, leading to inaccurate estimates (Montgomery et al., 2018).

However, this principle does not apply to questions addressing respondents' socio-demographic characteristics, such as age, gender, and education level. These variables are unlikely to be influenced by treatment exposure, unlike general attitudes or views on political issues. Therefore, researchers can safely place socio-demographic questions after the experimental section of the survey and use these variables in their analysis without introducing post-treatment bias.

From this principle, it follows that placing the experimental section toward the end of the survey, just before the outcome variable and socio-demographic questions, is optimal. Researchers can use the responses to preceding questions in their analysis. Yet, positioning many questions before the experimental section has a drawback: these questions might predispose respondents to a specific mindset, affecting their responses to the experiment. For instance, if a study aims to measure the effect of negative information about a racial group on civic engagement, including questions on racial resentment at the start could "prime" respondents to think about race, potentially affecting their reactions to the experiment. This does not compromise the estimates' validity but raises concerns about their external validity (Klar et al., 2020). It is unclear whether the same effects would be observed without predisposing respondents to a racial mindset from the outset.

Therefore, my advice to researchers is to be as parsimonious as possible in designing their survey questionnaires. They should include only the attention checks and questions that they know they will use for their heterogeneous analysis before the experimental section. After this section, they should add only the manipulation check, the designated outcome variables, and socio-demographic questions. However, I acknowledge that researchers do not always have complete control over their questionnaires. In such cases, I recommend transparency in their report of the results about any questions placed before the experimental section that could have primed respondents, and to refrain from conducting heterogeneous analyses or weighting the data based on responses to post-experimental questions, except for socio-demographic variables.

OTHER EXPERIMENTS IN POLITICAL SCIENCE

As I argue in the pages above, survey experiments have become the method of choice for experimental political scientists. It is, however, certainly not the only one. In this section, I will briefly discuss how other types of experiments, namely field and lab experiments, are used in political science. Given that this chapter is focused on survey experiments, I will not go into much detail. I will, however, provide some references for the researchers interested in those other methods.

Field Experiments

Historically, political science has been a pioneer discipline of the social sciences when it comes to field experiments. One of the first such experiments, where treatments are assigned by the researcher to a random sample of subjects in real life, and their effects are also observed in real life, was the Get-Out-To-Vote experiment conducted by Gosnell in Chicago in the 1920s (Gosnell, 1926). In this experiment, Gosnell tested the causal effect of an intervention meant to increase turnout (outcome Y), namely door-to-door canvassing where people are reminded about the upcoming election (treatment X). Since then, there have been thousands of field experiments of this genre, as well as others, in political science. Field experiments are particularly useful for estimating the causal effect of policy interventions in real life. However, they generally do not allow researchers to elicit the mechanism behind this effect, and their theoretical contribution to scholarship can therefore be somewhat limited. For the readers interested, I recommend the textbook of Gerber and Green (2012), which covers all the practical steps. They should also read the chapter by Samek and Sadoff in the present volume, as most of the topics discussed there, as well as the tips provided by the author, also apply to field experiments in political science (see Chapter 2).

Note, however, that a special class of field experiments has gained some popularity in political science over the last decades. These are audit experiments with elected officials where the researcher sends an email or a letter to politicians (elected or not) on behalf of a citizen asking for some information or service. They then observe whether these politicians respond and what the content of these responses is, typically conditional on some random treatments in the content of the letter, such as the race of the sender, as indicated by their name (Butler and Broockman, 2011), or their partisan preferences (Bol et al., 2021). These outcomes are then considered as measures of how responsive the contacted politicians are and how much they care about their constituents. Although promising, these audit experiments pose some important ethical challenges, notably due to the use of deception. For the reader interested in the topic, I recommend the review article of Grose (2014), as well as the special issue edited by John and Foos (2022) that specifically touches upon the ethical aspects of this method. Among other safety tips, they, for example, recommend asking real constituents to send a standardized email drafted by the researcher. This strategy ensures that the experiment does not involve any deception, but is also time-consuming as it requires the researcher to find a number of constituents willing to participate in the study and brief them accordingly.

Lab Experiments

There is also a long tradition of lab experiments in political science. In these experiments, the researcher reproduces political interactions between subjects, often students, in the controlled environment of a laboratory to evaluate the effect of randomized treatments on individual and group behaviors. Such experiments have been inspired by the pioneering work in behavioral economics as well as experimental psychology, and then applied to political topics such as voting or political behavior more generally (Bol, 2019). For example, Duffy and Tavits (2008) reproduce elections in a lab where subjects vote together to determine which of two options, A or B, wins. They then vary the chances with which each individual's vote affects the final result to evaluate how much the prospect of being pivotal affects the propensity of subjects to vote or to abstain.

Lab experiments remain an important method in political science because they allow researchers to fully control the experimental process and collect precise information on each subject. Thus, eliciting theoretical mechanisms, their use has faded somewhat in the last decade. This decline is likely due to the development of new technologies that enable researchers to conduct similar experiments on larger and more diverse samples, notably online. In a discipline focused on context-dependent phenomena (see above), lab experiments conducted with a student population are often seen as limited in their external validity, as results are not necessarily generalizable to a large set of people and countries (Bol, 2024).

A common strategy to address this issue involves complementing lab experiments with other methods, typically those based on surveys. The rationale is that if similar patterns emerge in both the small, convenient samples from the lab and the large, representative survey samples, it reinforces confidence in the generalizations. For example, Martinangeli et al. (2024) applied this method to study the effects of the quality of political institutions on people's propensity to trust each other. To do this, they first organized a lab experiment involving a series of trust games between subjects. The twist was that one of the subjects was chosen as the administrator of the games, who could decide to keep (some of) the money at play for their personal benefit. They found that the presence of a mechanism preventing administrators from stealing money from others, i.e., the treatment, increased the likelihood of subjects trusting and cooperating with others. After presenting their lab experiment results, they examined a large cross-country survey dataset for a correlation between respondents' trust in their co-nationals and the prevalence of corrupt behaviors among public officials in their country. They found a positive association, which corroborates their lab experimental results, thereby reinforcing their generalizability beyond the lab and in countries other than where their study was conducted.

For the readers interested in the way lab experiments are used in political science, I recommend the review paper that I wrote on the topic (Bol, 2019), as well as the textbook of Morton and Williams (2010) and the chapter by Niederle in the present volume (Chapter 1). Here again, the topics discussed as well as the tips given by the author largely apply to the way lab experiments are conducted in political science.

CONCLUSION

In this chapter, I have discussed various ways in which experimental methods are utilized in political science, focusing on survey experiments, which are now the most prevalent of these methods. After defining the contours of survey experiments and their various sub-types, I have provided practical guidelines and tips for researchers interested in using them. Toward the end, I also briefly discussed how other experimental methods, both in the lab and in the field, are used in political science. In this conclusion, I will discuss what I consider to be the two most important challenges faced by political scientists who use survey experiments.

A primary challenge lies in the representativeness of the samples. As mentioned earlier, a key reason why political scientists use survey experiments is that they allow the estimation of average treatment effects over a sample of respondents representative of a country's population. This is crucial for a discipline that studies context-dependent phenomena. However, the practice of relying on polling companies to provide online samples has undermined sample quality. As explained earlier, these companies use socio-demographic quotas to ensure

that each social group is represented in the sample proportionally to their actual size in the population. However, just because a sample is representative in terms of socio-demographic characteristics does not mean it is representative of other variables.

Ultimately, polling companies can only send questionnaires to individuals for whom they have contact information. Despite efforts to constitute a panel of potential respondents as diverse as possible, it is reasonable to think that they cannot reach every population stratum. This might include individuals who are rarely online or are socially secluded. Moreover, these individuals must agree to fill out surveys for the company. It is likely that some are too concerned about their privacy or distrustful of the corporate world to participate. In other words, while polling companies can provide samples representative of the population in observable socio-demographics, these samples might not represent the population in non-observable characteristics and likely have some blind spots.

There is little researchers can do to address this issue, except for contracting reputable polling companies, such as Understanding America Study (see <https://uasdata.usc.edu/index.php>), which uses census data and postal correspondence to maximize the representativeness of their sample. However, I encourage the political science community not to draw a definitive line between experiments conducted with samples made representative through socio-demographic quotas and those relying on purely convenient samples, such as those composed of students. The former is, of course, closer to the actual population than the latter, but the difference is more a matter of degree than of kind.

Second, another concern with survey experiments relates to the ethical treatment of respondents. Although much has been written about designing treatments that do not deceive or harm respondents, for example, by causing undue stress or anxiety (e.g., Desposato, 2015), the issue of their monetary compensation is often overlooked. As mentioned earlier, polling companies generally offer a small amount of money to persuade individuals to participate in a survey. However, they typically offer only a few dollars or even cents per survey. By using these companies, researchers may inadvertently contribute to exploiting individuals and creating precarious job conditions.

Here again, there is little researchers can do to address this issue, except for putting pressure on polling companies to increase their monetary incentives, for example, by boycotting those that pay the least. However, I encourage the political science community, and beyond, to be more considerate of the individuals they study and their time. Survey experiments are a powerful tool with much potential to advance our knowledge of how individuals approach politics, but we should collectively strive to conduct studies that follow the best practices of the discipline, both to avoid wasting the time and energy of our human subjects and as a sign of respect for these subjects.

NOTES

1. Most survey platforms include a tool to randomize questions or widgets in a questionnaire. This tool ensures that, *ex ante*, each respondent has an equal probability of being assigned to either the control group or any of the treatment groups. A common method to evaluate whether this randomization was done properly is to conduct an *ex-post* balance test, which involves regressing the treatment status of the respondents on answers to questions asked before the experimental part of the survey and/or answers that are not

- supposed to be affected by the treatments, such as socio-demographics. If the randomization is correct, none of these covariates should be associated with the treatment status.
2. Note that political scientists use the same estimation techniques as other experimentalists: they favor OLS regressions with the treatment status as the sole covariate. Note, however, that sometimes this estimation is complemented by another one where the observations of the control and treatment groups are re-weighted to make them perfectly balanced across socio-demographic variables with the use of a method called “entropy balancing” (Hainmueller, 2012). Even though randomization should, in principle, ensure some balance between the groups in this respect, there can be, by chance, some small discrepancies. Thanks to entropy balancing, these discrepancies disappear, and the estimates are more precise.
 3. For another example of research that leverages a list experiment to estimate attitudes towards the LGBTQI+ population, see Coffman et al. (2016).
 4. If the researcher can predict who among respondents is a complier based on covariates that cannot be affected by the treatment status, like the respondents’ socio-demographic characteristics, they can use these covariates to remove non-compliers from both the treatment and control groups or use them as instrumental variables in a two-stage regression analysis. This would generate accurate and precise estimates of the average treatment effect.

REFERENCES

- Adida, Claire L., Adeline Lo, and Melina R. Platas. 2018. “Perspective Taking Can Promote Short-Term Inclusionary Behavior Toward Syrian Refugees.” *Proceedings of the National Academy of Sciences*, 115 (38), 9521–9526.
- Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. “The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics.” *Journal of Economic Perspectives*, 24 (2), 3–30.
- Arel-Bundock, Vincent, Ryan C. Briggs, Hristos Doucouliagos, Marco Mendoza Aviña, and Tom D. Stanley. 2024. “Quantitative Political Science Research Is Greatly Underpowered.” I4R Discussion Paper Series, The Institute for Replication (I4R). www.econstor.eu/handle/10419/265531.
- Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto. 2018. “The Number of Choice Tasks and Survey Satisficing in Conjoint Experiments.” *Political Analysis*, 26 (1), 112–119.
- Blair, Graeme, and Kosuke Imai. 2012. “Statistical Analysis of List Experiments.” *Political Analysis*, 20 (1), 47–77.
- Blais, André, Damien Bol, Shaun Bowler, David M. Farrell, Annika Fredén, Martial Foucault, Emmanuel Heisbourg, Romain Lachat, Ignacio Lago, Peter J. Loewen, Miroslav Nemčok, Jean-Benoit Pilet, and Carolina Plescia. 2023. “What Kind of Electoral Outcome Do People Think Is Good for Democracy?” *Political Studies*, 71 (4), 1068–1089.
- Bol, Damien. 2019. “Putting Politics in the Lab: A Review of Lab Experiments in Political Science.” *Government and Opposition*, 54 (1), 167–190.
- Bol, Damien. 2024. “Laboratory Experiments in Political Science.” In Bereket Kebede (ed.) *Encyclopedia of Experimental Social Science*. Edward Elgar Publishing.
- Bol, Damien, André Blais, and Jean-François Laslier. 2018. “A Mixed-Utility Theory of Vote Choice Regret.” *Public Choice*, 176 (3), 461–478.
- Bol, Damien, Thomas Gschwend, Thomas Zittel, and Steffen Zittlau. 2021. “The Importance of Voter Signals for the Responsiveness of Legislators: A Field Experiment.” *European Journal of Political Research*, 60 (2), 455–473.
- Butler, Daniel M., and David E. Broockman. 2011. “Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators.” *American Journal of Political Science*, 55 (3), 463–477.

- Camarer, Colin F., and Robin M. Hogarth. 1999. "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework." *Journal of Risk and Uncertainty*, 19, 7–42.
- Coffman, Katherine B., Lucas C. Coffman, and Keith M. Marzilli Ericson. 2016. "The Size of the LGBT Population and the Magnitude of Antigay Sentiment Are Substantially Underestimated." *Management Science*, 63 (10), 3168–3186.
- Desposato, Scott (ed.). 2015. *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*. Routledge.
- Druckman, James N., and Donald P. Green. 2021. *Advances in Experimental Political Science*. Cambridge University Press.
- Duffy, John, and Margit Tavits. 2008. "Beliefs and Voting Decisions: A Test of the Pivotal Voter Model." *American Journal of Political Science* 52(3): 603–618.
- Egami, Naoki, and Erin Hartman. 2023. "Elements of External Validity: Framework, Design, and Analysis." *American Political Science Review*, 117 (3), 1070–1088.
- Esaiasson, Peter, and Patrik Öhberg. 2020. "The Moment You Decide, You Divide: How Politicians Assess Procedural Fairness." *European Journal of Political Research*, 59 (3), 714–730.
- Feddersen, Timothy, Sean Gailmard, and Alvaro Sandroni. 2009. "Moral Bias in Large Elections: Theory and Experimental Evidence." *American Political Science Review*, 103 (2), 175–192.
- Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. W. W. Norton.
- Gosnell, Harold F. 1926. "An Experiment in the Stimulation of Voting." *American Political Science Review*, 20 (4), 869–874.
- Grose, Christian R. 2014. "Field Experimental Work on Political Institutions." *Annual Review of Political Science*, 17, 355–370.
- Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis*, 20 (1), 25–46.
- Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis*, 22 (1), 1–30.
- John, Peter, and Florian Foos. 2022. "Introduction to Symposium: Experiments with Politicians: Ethics, Power, and the Boundaries of Political Science." *Political Studies Review*, 20 (2), 169–174.
- Kam, Cindy D., Jennifer R. Wilking, and Elizabeth Zechmeister. 2007. "Beyond the Narrow Data Base: Another Convenience Sample for Experimental Research." *Political Behavior*, 29 (4), 415–440.
- Kane, John V. 2024. "More Than Meets the ITT: A Guide for Anticipating and Investigating Nonsignificant Results in Survey Experiments." *Journal of Experimental Political Science*. <https://doi.org/10.1017/XPS.2024.1>
- Karp, Jeffrey A., and David Brockington. 2005. "Social Desirability and Response Validity: A Comparative Analysis of Overreporting Voter Turnout in Five Countries." *Journal of Politics*, 67 (3), 825–840.
- Klar, Samara, Thomas Leeper, and Joshua Robison. 2020. "Studying Identities with Experiments: Weighing the Risk of Posttreatment Bias Against Priming Effects." *Journal of Experimental Political Science*, 7 (1), 56–60.
- Kuklinski, James H., Paul M. Sniderman, Kathleen Knight, Thomas Piazza, Philipp E. Tetlock, Gordon R. Lawrence, and Barbara Mellers. 1997. "Racial Prejudice and Attitudes toward Affirmative Action." *American Journal of Political Science*, 41 (2), 402–419.
- Lazarsfeld, Paul F., Bernard Berelson, and Hazel Gaudet. 1948. *The People's Choice*. Columbia University Press.
- Martinangeli, Andrea F.M., Marina Povitkina, Sverker Jagers, and Bo Rothstein. 2024. "Institutional Quality Causes Generalized Trust: Experimental Evidence on Trusting under the Shadow of Doubt." *American Journal of Political Science*, 68 (3), 972–987.
- Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. "How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It." *American Journal of Political Science*, 62 (3), 760–775.
- Morin-Chassé, Alexandre, Damien Bol, Laura B. Stephenson, and Simon Labbé St-Vincent. 2017. "How to Survey About Electoral Turnout? The Efficacy of the Face-Saving Response Items in 19 Different Contexts." *Political Science Research and Methods*, 5 (3), 575–584.

- Morton, Rebecca B., and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. Cambridge University Press.
- Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences*, 115 (11), 2600–2606.
- Pilet, Jean-Benoit, Damien Bol, Davide Vittori, and Emilien Paulis. 2023. "Public Support for Deliberative Citizens' Assemblies Selected through Sortition: Evidence from 15 Countries." *European Journal of Political Research*, 62 (3), 873–902.
- Samuels, David, and Cesar Zucco. 2014. "The Power of Partisanship in Brazil: Evidence from Survey Experiments." *American Journal of Political Science*, 58 (1), 212–225.
- Sheffer, Lior, Peter John Loewen, Stuart Soroka, Stefaan Walgrave, and Tamir Sheafer. 2018. "Nonrepresentative Representatives: An Experimental Study of the Decision Making of Elected Politicians." *American Political Science Review*, 112 (2), 302–321.
- Teele, Dawn Langan, Joshua Kalla, and Frances Rosenbluth. 2018. "The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics." *American Political Science Review*, 112 (3), 525–541.
- Thomas, Kathrin. 2024. "The Advent of Survey Experiments in Politics and International Relations." *Government and Opposition*, 59 (1), 297–320.
- Torreblanca, Carolina, William Dinneen, Guy Grossman, and Yiqing Xu (2026). "The Credibility Revolution in Political Science." Working Paper available at: <https://arxiv.org/abs/2601.11542>
- Villeval, Marie-Claire. 2007. "Experimental Economics: Contributions, Recent Developments, and New Challenges." *French Politics*, 5 (2), 178–186.
- Westwood, Sean J. 2025. "The Potential Existential Threat of Large Language Models to Online Survey Research." *Proceeding of the National Academy of Sciences* 122(47): e2518075122.