Table of Contents

# Abstract

This guide discusses techniques for using randomization to create experiments within the text of a survey (i.e. survey experiments). These survey experiments are distinct from studies that use surveys to gather information related to an experiment that occurs outside of the survey. The guide distinguishes between survey experiments that are used mainly for measuring sensitive attitudes, like list experiments, and those that are mainly used to learn about causal effects, like conjoint experiments. Survey experiments for measurement attempt to ensure honest responses to sensitive questions by providing anonymity to respondents. Survey experiments for causal identification randomize images and text to learn how the image or text influences respondents. Both types of survey experiments face challenges, such as respondents not perceiving anonymity or not interpreting images and text as the researcher intended. New experimental techniques seek to address these challenges.

# 1 What is a Survey Experiment

A survey experiment is an experiment conducted within a survey. In an experiment, a researcher randomly assigns participants to at least two experimental conditions. The researcher then treats each condition differently. Due to random assignment, the researcher can assume that the only difference between conditions is the difference in treatment. For example, a medical experiment may learn about the effect of a pill by creating two experimental conditions and giving the pill to participants in only one condition. In a survey experiment, the randomization and treatment occur within a survey questionnaire.

There are two types of survey experiments. One type is used to measure sensitive attitudes or behaviors and the other is used to learn about causal relationships. By sensitive attitudes and behaviors, we mean any attitude or behavior that the respondent does not want to be publicly associated with. Many respondents, for example, do not want to be associated with racism or illegal behaviors.

Survey experiments for measurement attempt to provide respondents with anonymity so that they can express potentially sensitive attitudes without being identified as holding the sensitive attitude. These measurement survey experiments are alternatives to asking direct questions when direct questions are likely subject to response biases (i.e. when the respondents are likely to lie). These indirect measures are especially useful in contexts where direct questions may be dangerous for survey respondents and enumerators (Bullock, Imai, and Shapiro 2011).

Survey experiments to measure causal relationships are just like any other experiment, except the experimental intervention and outcome measurement occur within the context of a survey. Researchers randomly assign respondents to one or more experimental conditions or a control group, administer treatments, and measure outcomes. As their experimental manipulation, survey experiments tend to present a randomized stimulus (such as text, short videos, audio clips, or other stimuli that can easily fit into a survey format) in the context of a survey question or use variation in question wording and question ordering. They often administer these randomized manipulations by using survey questionnaire software or by providing different physical survey forms to enumerators.

# 2 Survey Experiments to Measure Sensitive Topics: List Experiment

List experiments (also known as the item-count technique) measure a sensitive attitude or behavior when the researcher expects the respondent to lie if asked a direct question about the attitude or behavior. For example, respondents may be reluctant to admit racial animus (Kuklinski et al. 1997) or illegal behavior (Garcia-Sanchez and Quierolo, n.d.), even after being assured of the survey’s anonymity.

In a list experiment the researcher randomly assigns respondents to a control or treatment condition. The control condition presents respondents with a list of items; the treatment condition presents respondents with the same list plus a treatment item measuring the attitude or behavior of interest. Respondents are then asked how many of those items apply to them. The average difference between the treatment and control condition represents the percentage of respondents for whom the treatment item applies. A list experiment does not tell the researcher about the attitude of any individual respondent, but it tells her about the prevalence of the sensitive attitude in her survey population. Answers to this question are anonymous because the respondent’s attitude towards each item cannot be determined unless the respondent answers that all or none of the items apply to them. Respondents for whom the treatment item applies can increase their response by one without revealing their attitude.

For example, Kuklinski et al. (1997) studied racial animus with a survey list experiment. They told respondents:

Now I am going to read you three things that sometimes make people angry or upset. After I read all three, just tell me HOW MANY of them upset you. I don’t want to know which ones, just HOW MANY. (1) the federal government increasing the tax on gasoline (2) professional athletes getting million-dollar contracts (3) large corporations polluting the environment *(4) a black family moving in next door*

The 4th item was withheld from the control condition.

For that paper’s population of interest, the mean number of items chosen in the treatment group was 2.37, compared to 1.95 in the control. The difference of 0.42 between treatment and control indicates that 42% of respondents would be upset by a black family moving in next door. **Pitfalls of list experiments**

List experiments are *vulnerable to satisficing*. Satisficing occurs when respondents put in minimal effort to understand and answer a survey question (Krosnick 1991; Simon and March 2006). In a list experiment, satisficing manifests when respondents do not count the number of items that apply to them, instead answering with a number of items that seems reasonable (Kramon and Weghorst 2012; Schwarz 1999).

Respondents may *perceive a lack of anonymity*. Despite the anonymity provided by a list experiment, respondents may still worry that their response reflects their attitudes about the sensitive item. When respondents worry about a lack of anonymity, they may increase or decrease their response to portray themselves in the best light possible, rather than answer honestly (Leary and Kowalski 1990). For example, the addition of a treatment item about race can *decrease* the number of items that respondents report because being associated with “three of the four [list items] may be interpreted as a 75% chance that they are racist” (Zigerell 2011, 544).

The lack of anonymity is most obvious when all or none of the list items apply to the respondent. Researchers can reduce this possibility by using *uncorrelated or negatively correlated control items* that are unlikely to apply to one respondent. In the Kuklinski et al. (1997) example above, the type of person who is upset by pollution is unlikely to also be upset by a gasoline tax. Negatively correlated items also reduce likelihood that respondent will satisfice because negatively correlated items are unlikely to be interpreted as a scale measuring one concept. The control items should also fit with the treatment item in some way so that the treatment item does not jump out to respondents as the real item of interest to researchers. **Variants/Modifications**

*Double list experiments* help overcome some pitfalls of single list experiments (Glynn 2013; Droitcour et al. 2004). In a double list experiment, the treatment item is randomly selected to appear on either the first or the second control list, so that some respondents see it on the first list and some respondents see it on the second. If researchers observe the same treatment effect on both lists, there is less risk that the effect depends on a particular control list or on how respondents interpret the list. The double list experiment is also more statistically efficient than a single list experiment (Glynn 2013).

*Placebo-controlled list experiments* ensure that the difference in responses to the treatment and control lists is due to the treatment item and not due to the treatment list having more items than the control list. A placebo-controlled list experiment uses an additional item as a placebo on the control list; unlike the additional item on the treatment list, the additional item on the control list is something innocuous that would not apply to any respondent. The placebo item ensures that the difference between the two lists is due to the treatment item, not the presence of an additional item (Riambau and Ostwald 2019).

*Visual aids* also help reduce satisficing and ensure that respondents follow the instruction to count list items instead of satisfice. If enumerators can carry a laminated copy of the list and a dry erase marker, respondents can check off items on the list to get an exact count and erase it before handing it back to the enumerator (Kramon and Weghorst 2012, 2019).

# 3 Survey Experiments to Measure Sensitive Topics: Randomized Response

The randomized response technique is also used to measure a sensitive attitude or behavior when the researcher expects the respondent to lie if asked a direct question (Warner 1965; Boruch 1971; Gingerich 2015, 2010).

In the most common version of the randomized response technique, respondents are directly asked a yes or no question about a sensitive topic. The respondent is also given some randomization device, like a coin or die. The respondent is told to answer the direct question when the randomization device takes on a certain value (tails) or to say “yes” when the randomization device takes a different value (heads). Researchers assume that respondents will believe their anonymity is protected because the researcher cannot know whether a “yes” resulted from agreement with the sensitive item or the randomization device.

For example, Blair, Imai, and Zhou (2015) studied support for militants in Afghanistan with the randomized response technique. They gave respondents a die and had the respondent practice throwing it. They then told respondents:

For this question, I want you to answer yes or no. But I want you to consider the number of your dice throw. If 1 shows on the dice, tell me no. If 6 shows, tell me yes. But if another number, like 2 or 3 or 4 or 5 shows, tell me your opinion about the question that I will ask you after you throw the dice. [ENUMERATOR TURN AWAY FROM THE RESPONDENT] Now throw the dice so that I cannot see what comes out. Please do not forget the number that comes out. [ENUMERATOR WAIT TO TURN AROUND UNTIL RESPONDENT SAYS YES TO]: Have you thrown the dice? Have you picked it up? Now, during the height of the conflict in 2007 and 2008, did you know any militants, like a family member, a friend, or someone you talked to on a regular basis? Please, before you answer, take note of the number you rolled on the dice.

In expectation, 1/6th of respondents answer “yes” due to the die throw. The researcher can thus determine what percentage of respondents engaged in the sensitive behavior. **Pitfalls of the randomized response technique**

*Some versions are complicated*. Even the common version described above, valued in part for its simplicity, requires respondents to use some randomization device and remember the outcome of the randomization device. Other versions use more complicated techniques to ensure anonymity; these versions may be difficult both for the respondent and the enumerator (Blair, Imai, and Zhou 2015; Gingerich 2010). It is possible that some respondents do not understand the instructions and some enumerators do not implement the randomized response technique properly.

Respondents may *perceive a lack of anonymity*. As was true for list experiments, respondents may not feel that their answers to randomized response questions are truly anonymous. If a respondent answers “yes”, the answer could have been dictated by the randomization device, but it could also signal agreement with the sensitive item (Edgell, Himmelfarb, and Duchan 1982; Yu, Tian, and Tang 2008). Thus, answering “yes” is not unequivocally protected by the design. Edgell, Himmelfarb, and Duchan (1982) surreptitiously set the randomization device to always dictate “yes” or “no” for specific questions and observed as high as 26% of respondents say “no” even when the randomization device dictated they say “yes”. **Variants/modifications**

The *repeated randomized response technique* helps researchers identify respondents who lie on randomized response questions (Azfar and Murrell 2009). The repeated technique asks a series of randomized response questions with sensitive and non-sensitive items. The probability of the randomization device dictating that the respondent should answer “no” for all of the *sensitive* items is very low. The technique thus allows researchers to identify and remove from analysis the respondents who are likely saying “no” even when their coin flip dictates they say “yes”. Researchers can also determine if certain questions induce widespread lying if the “yes” rate for that question is lower than the randomization device would dictate. The repeated randomized response technique, however, may be impractical to include on a large survey.

The *Crosswise model* modifies the randomized response technique so that respondents have no incentive to answer “yes” or “no” (Yu, Tian, and Tang 2008; Jann, Jerke, and Krumpal 2011). In the Crosswise model, respondents are presented with two statements, one sensitive statement and one non-sensitive statement for which the population mean is known. The respondent is asked to say if (a) neither or both statements are true, or (b) one statement is true. Unlike a typical randomized response question, where individuals who agree with the sensitive statement only occupy the “yes” group, people who agree with the sensitive statement could occupy either group using the Crosswise model. Since being in category (a) and (b) are equally uninformative about the respondent’s agreement with the sensitive statement, the Crosswise model removes a respondent’s incentive to lie. The Crosswise model can be used any time researchers know the population mean of a non-sensitive statement, such as “My mother was born in April.”

# 4 Survey Experiments to Measure Sensitive Topics: Priming Experiment

List experiments and randomized response techniques do not uncover implicit attitudes, but many sensitive topics appear so sensitive that an individual’s conscious, explicit attitudes may differ from their implicit attitudes (Greenwald and Banaji 1995). Even many nonsensitive attitudes seem to be beyond an individual’s conscious awareness (Nisbett and Wilson 1977). Whereas techniques to measure explicit attitudes seek to provide respondents with anonymity, techniques to measure implicit attitudes seek to keep the respondent consciously unaware of the implicit attitude being measured. To do so, researchers often use priming experiments.

In a priming experiment, researchers expose respondents to a stimulus representing topic *X* in order to influence their response to a survey question about topic *Y*, without the respondent realizing that the researchers are interested in topic *X*. A control group is not exposed to the stimuli representing topic *X*, so the difference between the treatment group and control group is due to exposure to the treatment stimuli. Priming experiments work by directing respondents’ consciousness away from topic *X* and towards topic *Y* so that respondents do not consciously censor their feelings about topic *X* (Macrae et al. 1994; Schwarz and Clore 1983).

Priming experiments are a broad class and include any experiment that makes a sensitive topic salient in the mind of the respondent. One common method of priming is the use of images. For example, Brader, Valentino, and Suhay (2008) use images in a priming experiment to estimate the effect that race plays in opposition to immigration. The researchers show subjects a positive or negative news article about immigration paired with a picture of a European immigrant or an Hispanic immigrant. Subjects expressed negative attitudes about immigration when the negative news article is paired with the Hispanic immigrant picture but not in other conditions. The picture primes people to think about Hispanic immigrants, and thinking about Hispanic immigrants reduces support for immigration compared to thinking about European immigrants even though subjects do not consciously admit to bias.

Survey experiments for measurement and for causal identification overlap in priming experiments. Researchers can use them to measure implicit attitudes or to assess how the activation of implicit attitudes affects another outcome, like attitudes towards immigration. **Pitfalls of priming experiments**

*Priming experiments are difficult*. Priming attitudes experimentally is difficult because the researcher cannot be certain that the prime affects subjects as the researcher intended. A prime intended to induce fear, for example, may induce fear in some subjects and excitement in others. Priming sensitive attitudes is especially difficult because the researcher must prime a sensitive attitude without the respondent becoming aware that the researcher is interested in the sensitive attitude. If respondents realize what the priming experiment is about, the experiment fails because respondents will consciously censor their attitude, rather than passively allow their implicit attitude to influence their response (Macrae et al. 1994; Schwarz and Clore 1983). To prevent subjects from ascertaining the goal of the study, researchers try to hide the prime amid other, ostensibly more important, information.

*Priming experiments can suffer from confounding and lack of “information equivalence” between treatment groups* (Dafoe, Zhang, and Caughey 2018). The researchers may prime topic with the intent of learning about respondents’ implicit attitudes towards topic , but if topic is strongly linked with topic then the researcher will estimate the effect of and , not just . For example, priming a partisan group may also prime ideological and policy views associated with the partisan group (Nicholson 2011). A basic priming experiment cannot differentiate the effect of priming the partisan group from the effect of priming the ideological and policy views associated with the partisan group.

*Respondents may be pretreated before the experiment*. Individuals are exposed to stimuli that prime attitudes during their daily lives. News broadcasts prime people to think about issues covered on the news, and anti-racism protests prime people to think about racial issues. Even words seen immediately before answering survey questions influences responses to those survey questions (Norenzayan and Schwarz 1999). If subjects, before participating in the experiment, encounter the stimuli that the researcher wants to prime, there may be no difference between treatment and control groups because all subjects were “pretreated” with the prime, even subjects in the control group (Gaines, Kuklinski, and Quirk 2007; Druckman and Leeper 2012). If the issue being primed is already salient in the mind of the respondent, priming experiments fail. **Variants/modifications**

To ensure information equivalence and to reduce confounding the prime with an associated factor, researchers utilize priming experiments as part of *factorial experiments*. Factorial experiments vary multiple factors that may be linked in the minds of respondents. Nicholson (2011), for example, asked respondents about support for a policy. He varied both partisan endorsement and policy details to learn how partisan bias influenced respondents’ attitudes beyond any assumptions about the party’s policy positions. Factorial experiments are mainly used to determine causal relationships and are discussed in section 7.

# 5 Survey Experiments to Measure Sensitive Topics: Endorsement Experiments

Endorsement experiments measure sensitive attitudes towards an attitude object, like a political actor or a policy. They were first developed to study partisan bias (Cohen 2003; Kam 2005) but have since been used to measure support for militant groups (Bullock, Imai, and Shapiro 2011; Lyall, Blair, and Imai 2013). They have also been inverted to measure support for a policy rather than a political actor (Rosenfeld, Imai, and Shapiro 2016).

In a typical endorsement experiment, respondents are asked how much they support a policy. In the treatment condition, the policy is “endorsed” by a group that respondents would not consciously admit to influencing their opinion. In the control condition, the policy is not endorsed by any group. The average difference in support between the endorsed and unendorsed policy represents the change in support for the policy because of the endorsement.

Endorsement experiments can measure implicit attitudes or explicit attitudes. They measure implicit attitudes like a priming experiment if respondents do not realize the group’s endorsement is what the researcher is interested in. They measure explicit attitudes like a list experiment if respondents realize the group’s endorsement is what the researcher is interested in. Whereas list experiments hide the respondent’s opinion by pairing the sensitive item with non-sensitive control items, endorsement experiments hide the respondent’s opinion by pairing the sensitive item with a policy that could feasibly be responsible for the respondent’s attitude.

For example, Nicholson (2012) used an endorsement experiment to study partisan bias in the United States during the 2008 Presidential campaign. The researchers asked respondents about policies, varying whether the policy was unendorsed or endorsed by the Presidential candidates of the two main political parties, Barack Obama (Democrat) and John McCain (Republican). Respondents were told:

As you know, there has been a lot of talk about immigration reform policy in the news. One proposal [**backed by Barack Obama**/**backed by John McCain**] provided legal status and a path to legal citizenship for the approximately 12 million illegal immigrants currently residing in the United States. What is your view of this immigration reform policy?

The difference between the control condition and the Obama (McCain) condition for Democrats (Republicans) shows in-party bias. The difference between the control condition and the Obama (McCain) condition for Republicans (Democrats) shows out-party bias. **Pitfalls of endorsement experiments**

As with priming experiments, endorsement experiments suffer from *confounding and a lack of information equivalence*. Researchers cannot be certain if differential support for the policy is due to the endorsement or due to different substantive assumptions about the policy that respondents make as a result of the endorsement.

*Choosing a policy is difficult*. The value of the endorsement experiments depends largely on the characteristics of the policy being (or not being) endorsed. The chosen policy must not possess too much or too little support in the survey population, otherwise attitudes towards the policy will wipe out the effect of the group’s endorsement. Too much or too little support could also reduce perceived anonymity if respondents think that no one would support/oppose the policy unless they liked/disliked the endorsing group.

Endorsement experiments can have *low power to detect effects*, even relative to other survey experiments (Bullock, Imai, and Shapiro 2011). Some subset of subjects will be unaffected by the endorsement because they feel strongly about the policy, and that subset adds substantial noise to endorsement experiments. **Variants/modifications**

To overcome low power, Bullock, Imai, and Shapiro (2011) recommend using *multiple policy questions* that are on the same one-dimensional policy space. Multiple questions on one policy space allows the researcher to predict each respondent’s level of support for the policy if it were not endorsed by the group of interest. The researcher can thus model the noise caused by strong feelings towards the policy.

As with priming experiments, to ensure information equivalence and to reduce confounding factors, researchers use endorsement experiments as part of *factorial experiments* that vary the multiple factors that may be linked in the mind of respondents. Factorial experiments are mainly used to determine causal relationships and are discussed in section 7.

# 6 Limitations of Survey Experiments as a Measurement Technique

Survey experiments induce less bias than direct questions when measuring sensitive attitudes (Blair, Imai, and Zhou 2015; Rosenfeld, Imai, and Shapiro 2016; Lensvelt-Mulders et al. 2005). They are not a panacea, however, and researchers must still ask themselves several questions when using survey experiments to measure sensitive outcomes.

The first question is whether the researcher is interested in an explicit or implicit attitude. An explicit attitude is one the respondent is consciously aware of and can report; an implicit attitude is an automatic positive or negative evaluation of an attitude object that the respondent may not be aware of (see Nosek 2007 for a more thorough discussion). A list experiment, for example, may help uncover explicit racial animus, but it will not reveal implicit racial bias.

The next question is what conditions are necessary for a survey respondent to reveal their explicit attitudes. Survey experimental methods for sensitive explicit attitudes focus on ensuring anonymity. But is ensuring anonymity a sufficient condition to obtain honest answers to sensitive questions? In addition to anonymity, a further assumption must be made: respondents want to express their socially undesirable opinion in a way that evades social sanctions. If that assumption is not true, then anonymity is worth little (Diaz, Grady, and Kuklinski 2020).

Researchers also need to think about the numerous pitfalls of survey questions that measurement survey experiments do not solve. Survey experiments do not help researchers avoid question ordering effects or contamination from earlier questions in the survey. Nor do they do expose how respondents interpret the survey question or ensure information equivalence. All survey questions assume that the respondent interprets the question in the way intended by researchers; techniques to ensure anonymity may make that interpretation less likely by obfuscating the question’s purpose (Diaz, Grady, and Kuklinski 2020).

Lastly, researchers must also ask about measurement validity: how does one verify that a measure accurately represents a concept of interest? For some outcomes, such as voter turnout, researchers can compare their measure with population estimates (Rosenfeld, Imai, and Shapiro 2016). But for other outcomes, such as racism or the effect that political parties have on citizens’ policy preferences, there exists no population estimate with which to validate measures.

# 7 Survey Experiments to Determine a Causal Relationship: Vignette and Factorial Designs

Not all survey experiments share the goal of accurately measuring one concept of interest. Some survey experiments, like lab experiments, are interested in how an experimental manipulation impacts outcomes of interest. These survey experiments for causal inference randomize a treatment and then measure outcomes. When measuring outcomes, they may use techniques like list experiments.

One of the most common designs of survey experiments for causal inference are vignette and factorial designs (Auspurg and Hinz 2014; Sniderman et al. 1991). In a vignette/factorial experiment, the researcher provides the respondent with a hypothetical scenario to read, varying key components of the scenario. In a typical vignette, the researcher varies only one component of the scenario. In a typical factorial experiment, the researcher varies several components of the scenario.

Both vignette and factorial designs benefit from embedding the survey question in a concrete scenario so that they require little abstraction from the survey respondent. Their concrete nature can make them more interesting and easier to answer than typical survey questions, decreasing survey fatigue. They can also function as priming experiments if the concept of interest is embedded in other concepts.

As an example, Winters and Weitz-Shapiro (2013) uses factorial vignettes to learn if voters sanction corrupt politicians in Brazil. They posit that corruption could interact with competence, so the authors varied a Brazilian mayor’s corruption, competence, and political affiliation in the vignette. They tell respondents:

Imagine a person named Gabriel (Gabriela for female respondents), who is a person like you, living in a neighborhood like yours, but in a different city in Brazil. The mayor of Gabriel’s city is running for reelection in October. He is a member of the [**Partido dos Trabalhadores**/**Partido da Social Democracia Brasileira**]. In Gabriel’s city, it is well known that the mayor [**never takes bribes**/**frequently takes bribes**] when giving out government contracts. The mayor has completed [**few**/**many**/**omit the entire sentence**] public works projects during his term in office. In this city, the election for mayor is expected to be very close. In your opinion, what is the likelihood that Gabriel(a) will vote for this mayor in the next election: very likely, somewhat likely, unlikely, not at all likely?

This design allowed the authors to determine if and when corruption would be punished by voters. If respondents overlooked corruption when the mayor completed many public works, the interpretation is that corruption is acceptable if it gets the job done. If respondents overlooked corruption when the mayor was a copartisan, the interpretation is that voters ignore the corruption of their own. By varying several related aspects of the scenario, Winters and Weitz-Shapiro (2013) could isolate the conditions under which credible information about corruption would be punished by voters. **Pitfalls of vignette/factorial experiments**

The main pitfall of vignettes – a *lack of information equivalence* – is dealt with by factorial experiments. Researchers can randomize several aspects of the scenario, standardizing factors that could influence how the main treatment is perceived by respondents. *Some combinations of different factors may not be realistic*, however. Researchers must be sure that the various possible combinations of their factorial experiments seem credible to respondents.

*Statistical power is weak* when factorial experiments vary many confounding traits. The more traits being varied, the more experimental conditions, the fewer respondents in each experimental condition, and the greater likelihood of imbalance between treatment conditions.

In enumerated surveys, there is also the possibility that certain enumerators are more often assigned certain factorial conditions and that *enumerator effects could be mistaken for treatment effects* (Steiner, Atzmüller, and Su 2016). Imagine the Winters and Weitz-Shapiro (2013) study, which had six functional treatment groups and ~2,000 respondents. If the survey was enumerated by twenty survey enumerators, then, in expectation, each enumerator has only ~17 subjects in each treatment category. In reality, it is likely that certain enumerators will more often enumerate some conditions than others and differences due to enumerators could appear as treatment effects. **Variants/modifications**

Researchers can *block treatment by enumerator* so that enumerator effects cannot confound treatment effects (Steiner, Atzmüller, and Su 2016). Blocking and other techniques the authors propose should also increase statistical power by accounting for systematic error.

*Conjoint experiments* maintain many benefits of factorial experiments, but increase power by presenting multiple choice tasks instead of one. We discuss conjoint experiments in the next section.

# 8 Survey Experiments to Determine a Causal Relationship: Conjoint Experiments

Conjoint experiments (Hainmueller, Hopkins, and Yamamoto 2014; Green and Rao 1971) have gained popularity in response to the limits of vignette and factorial designs. Vignette and factorial designs suffer from a lack of information equivalence if they do not provide sufficient details about potentially confounding aspects of the scenario or a lack of statistical power if they do vary several traits. A typical conjoint experiment attempts to solve these problems by repeatedly asking respondents to choose between two distinct options and randomly varying the characteristics of those two options. Respondents may also be asked to rate each option on a scale. In both cases, respondents express their preferences towards a large number of pairings with randomized attributes, drastically increasing statistical power to detect effects of any one attribute relative to a one-shot factorial design.

Hainmueller, Hopkins, and Yamamoto (2014) demonstrate the use of conjoint experiments in a study about support for immigration. The authors showed respondents two immigrant profiles and asked (a) which immigrant the respondent would prefer be admitted to the Unites States and (b) how the respondent rated each immigrant on a scale from 1-7. The authors randomly varied nine attributes of the immigrants (gender, education, employment plans, job experience, profession, language skills, country of origin, reasons for applying, and prior trips to the United States), yielding thousands of unique immigrant profiles. This process was repeated five times so that each respondents saw and rated five pairs of immigrants. Through this procedure, the authors can assess how these randomly varied components influence support for the immigrant.

Respondents saw:

Through a conjoint experiment, researchers can learn about the average marginal effect of several aspects of a scenario, far more than would be feasible with a typical vignette or factorial design. Though researchers could include and vary an almost infinite number of characteristics, the best practice is to only vary traits that could confound the relationship between a primary explanatory variable and an outcome of interest, rather than varying any trait that might affect the outcome of interest (Diaz, Grady, and Kuklinski 2020). **Pitfalls of conjoint experiments**

The costs and benefits of conjoint experiments are still being actively researched. Thus far, two classes of critiques are common.

Results from conjoint experiments are *difficult to interpret*. Results of conjoint experiments’ target estimand, the Average Marginal Component Effect (AMCE), can “indicate the opposite of the true preference of the majority” (Abramson, Koçak, and Magazinnik 2019, 1). Other researchers have noted that AMCE’s depend on the reference category and are not comparable across survey subgroups (Leeper, Hobolt, and Tilley 2020). Bansak et al. (2020) provides guidance on how to interpret conjoint results and argues that AMCE’s do represent quantities of interest to empirical scholars.

Conjoint experiments also create *unrealistic combinations* and those unrealistic combinations lead to effect estimates that are not representative of the real world (Incerti 2020). Similarly, the large amount of information provided by conjoint experiments could misrepresent how individuals generally process information they encounter in the world (Hainmueller, Hopkins, and Yamamoto 2014). The large amount of information and demand on respondents has also led to concerns about satisficing, though Bansak et al. (2018) and Bansak et al. (2019) suggest satisficing is not a major concern for conjoint experiments.

Other potential pitfalls can occur if the researcher varies too many characteristics. More randomly varied characteristics means a large number of potential hypothesis tests. The necessity of applying multiple hypothesis corrections to the vast number of potential hypothesis tests could decrease statistical power to detect specific effects, especially if researchers are interested in interactions between traits being varied.

# 9 Limitations of Survey Experiments for Causal Identification

Survey experiments to determine causal relationships have the same benefits and drawbacks as other experiments, as well as benefits and drawbacks that derive from the survey context. The biggest three drawbacks generally applicable to survey experiments are confounding, information equivalence, and pre-treatment contamination (Diaz, Grady, and Kuklinski 2020). Researchers should think about these factors when designing and interpreting results from survey experiments.

*Confounding*: Any experimental intervention *A* that is meant to trigger mental construct *M* could also trigger mental construct *C*. If *C* is not varied in the experimental design, researchers cannot determine whether *M*, *C*, or a combination of *M* and *C* affect outcomes of interest.

*Information Equivalence*: Any experimental intervention *A* can be interpreted differently by different respondents, effectively giving each respondent a different treatment (Dafoe, Zhang, and Caughey 2018). When these interpretations vary systematically by treatment condition, those conditions are not information equivalent and researchers cannot know that their treatment caused the observed effect.

*Pre-treatment contamination*: Respondents may encounter the treatment outside of the experiment, causing similar outcomes in the control group and treatment group even if the treatment affects outcomes (Gaines, Kuklinski, and Quirk 2007).

# 10 Considerations when using Survey Experiments

Survey experiments can be an effective tool for researchers to measure sensitive attitudes and learn about causal relationships. They are cost-effective, can be done quickly and iteratively, can be included on mass online surveys because they do not require in-person contact to implement. This means that a researcher can plan a sequence of online survey experiments, changing the intervention and measured outcomes from one experiment to the next to learn about the mechanisms behind the treatment effect very quickly (Sniderman 2018).

For survey experiments as a measurement technique, the researcher first has to assess if the attitude of interest is explicit (consciously known to the respondent) or implicit (not consciously known to the respondent). If the researcher believes the respondent knows her own attitude but does not want to be identified with it, the researcher should make it possible for the respondent to express that attitude without the researcher knowing that attitude. List experiments, randomized response techniques, and endorsement experiments can help accomplish this task. If the researcher believes the respondent does not know her own attitude, the researcher should make that attitude salient through priming and then ask a question that should be implicitly affected by the prime.

There may be cases where survey experiments are not the best tool for measuring sensitive attitudes. As an alternative to survey experiments to measure explicit attitudes, researchers can use techniques like the Bogus Pipeline (Jones and Sigall 1971) or phrase questions about a sensitive topic so that they are not considered socially undesirable (Kinder and Sears 1981). As an alternative to survey experiments to measure implicit attitudes, researchers can use measures like the Implicit Association Test (IAT) (Greenwald, McGhee, and Schwartz 1998) and physiological measures like skin conductance (Rankin and Campbell 1955; Figner, Murphy, and others 2011). These measures are beyond conscious control of the respondent. Many of these alternative measures are not currently flexible enough to be included on a mass survey, but technology, like heart-rate monitoring watches and other phone sensors, may soon make biometric outcomes measurable in mass surveys.

For all types of survey experiments, researchers should worry about the same issues that hamper other experiments: confounding, information equivalence, and pre-treatment contamination. To deal with confounding and information equivalence, researchers can design the experiment to manipulate characteristics that might confound the treatment. To account for pre-treatment, researchers can think about the everyday context of research subjects and assess whether all or a subset of respondents may already be treated before beginning the experiment. If only a subset will be affected, the researcher can block the experiment on that subset.

Survey experiments for measurement and survey experiments for estimating causal relationships are not binary categories, and the two types of survey experiments can overlap. Priming experiments, for example, can measure implicit attitudes and assess the effect of the prime on other outcomes of interest. Vignette or conjoint experiments can effectively measure a sensitive attitude by priming the sensitive attitude and providing lots of other information to distract respondents from the prime.

For more discussion of survey experiments, see:

* Mutz (2011) “Population-Based Survey Experiments.”
* Sniderman (2018) “Some Advances in the Design of Survey Experiments” in the Annual Review of Political Science.
* Lavrakas et al. (2019) “Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment.”
* Diaz, Grady, and Kuklinski (2020) “Survey Experiments and the Quest for Valid Interpretation” in the Sage Handbook of Research Methods in Political Science and International Relations.

# References

Abramson, Scott F, Korhan Koçak, and Asya Magazinnik. 2019. “What Do We Learn About Voter Preferences from Conjoint Experiments?” *Unpublished Manuscript. Https://Pdfs. Semanticscholar. Org/023a/24a7dfaddfce626d011596b187f26361ee86. Pdf*.

Auspurg, Katrin, and Thomas Hinz. 2014. *Factorial Survey Experiments*. Vol. 175. Sage Publications.

Azfar, Omar, and Peter Murrell. 2009. “Identifying Reticent Respondents: Assessing the Quality of Survey Data on Corruption and Values.” *Economic Development and Cultural Change* 57 (2). The University of Chicago Press: 387–411.

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). Cambridge University Press: 112–19.

———. 2019. “Beyond the Breaking Point? Survey Satisficing in Conjoint Experiments.” *Political Science Research and Methods*. Cambridge University Press, 1–19.

———. 2020. “Using Conjoint Experiments to Analyze Elections: The Essential Role of the Average Marginal Component Effect (Amce).” *Available at SSRN*.

Blair, Graeme, Kosuke Imai, and Yang-Yang Zhou. 2015. “Design and Analysis of the Randomized Response Technique.” *Journal of the American Statistical Association* 110 (511). Taylor & Francis: 1304–19.

Boruch, Robert F. 1971. “Assuring Confidentiality of Responses in Social Research: A Note on Strategies.” *The American Sociologist*. JSTOR, 308–11.

Brader, Ted, Nicholas A Valentino, and Elizabeth Suhay. 2008. “What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat.” *American Journal of Political Science* 52 (4). Wiley Online Library: 959–78.

Bullock, Will, Kosuke Imai, and Jacob N Shapiro. 2011. “Statistical Analysis of Endorsement Experiments: Measuring Support for Militant Groups in Pakistan.” *Political Analysis* 19 (4). Cambridge University Press: 363–84.

Cohen, Geoffrey L. 2003. “Party over Policy: The Dominating Impact of Group Influence on Political Beliefs.” *Journal of Personality and Social Psychology* 85 (5). American Psychological Association: 808.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. “Information Equivalence in Survey Experiments.” *Political Analysis* 26 (4). Cambridge University Press: 399–416.

Diaz, Gustavo, Christopher Grady, and James H Kuklinski. 2020. “Survey Experiments and the Quest for Valid Interpretation.” In *SAGE Handbook of Research Methods in Political Science and International Relations*. SAGE Publishing.

Droitcour, Judith, Rachel A Caspar, Michael L Hubbard, Teresa L Parsley, Wendy Visscher, and Trena M Ezzati. 2004. “The Item Count Technique as a Method of Indirect Questioning: A Review of Its Development and a Case Study Application.” *Measurement Errors in Surveys*. Wiley Online Library, 185–210.

Druckman, James N, and Thomas J Leeper. 2012. “Learning More from Political Communication Experiments: Pretreatment and Its Effects.” *American Journal of Political Science* 56 (4). Wiley Online Library: 875–96.

Edgell, Stephen E, Samuel Himmelfarb, and Karen L Duchan. 1982. “Validity of Forced Responses in a Randomized Response Model.” *Sociological Methods & Research* 11 (1). Sage Publications: 89–100.

Figner, Bernd, Ryan O Murphy, and others. 2011. “Using Skin Conductance in Judgment and Decision Making Research.” *A Handbook of Process Tracing Methods for Decision Research*, 163–84.

Gaines, Brian J, James H Kuklinski, and Paul J Quirk. 2007. “The Logic of the Survey Experiment Reexamined.” *Political Analysis* 15 (1). Cambridge University Press: 1–20.

Garcia-Sanchez, Miguel, and Rosario Quierolo. n.d. “A Tale of Two Countries: Measuring Drug Consumption in Opposite Contexts - Evidence from Survey Experiments in Colombia and Uruguay.”

Gingerich, Daniel. 2015. “Randomized Response: Foundations and New Developments.” In *Newsletter of the Comparative Politics Organized Section of the American Political Science Association*.

Gingerich, Daniel W. 2010. “Understanding Off-the-Books Politics: Conducting Inference on the Determinants of Sensitive Behavior with Randomized Response Surveys.” *Political Analysis* 18 (3). Cambridge University Press: 349–80.

Glynn, Adam N. 2013. “What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment.” *Public Opinion Quarterly* 77 (S1). Oxford University Press UK: 159–72.

Green, Paul E, and Vithala R Rao. 1971. “Conjoint Measurement-for Quantifying Judgmental Data.” *Journal of Marketing Research* 8 (3). SAGE Publications Sage CA: Los Angeles, CA: 355–63.

Greenwald, Anthony G, and Mahzarin R Banaji. 1995. “Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes.” *Psychological Review* 102 (1). American Psychological Association: 4.

Greenwald, Anthony G, Debbie E McGhee, and Jordan LK Schwartz. 1998. “Measuring Individual Differences in Implicit Cognition: The Implicit Association Test.” *Journal of Personality and Social Psychology* 74 (6). American Psychological Association: 1464.

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). Cambridge University Press: 1–30.

Incerti, Trevor. 2020. “Corruption Information and Vote Share: A Meta-Analysis and Lessons for Experimental Design.” *American Political Science Review, Forthcoming*.

Jann, Ben, Julia Jerke, and Ivar Krumpal. 2011. “Asking Sensitive Questions Using the Crosswise Model: An Experimental Survey Measuring Plagiarism.” *Public Opinion Quarterly* 76 (1). Oxford University Press: 32–49.

Jones, Edward E, and Harold Sigall. 1971. “The Bogus Pipeline: A New Paradigm for Measuring Affect and Attitude.” *Psychological Bulletin* 76 (5). American Psychological Association: 349.

Kam, Cindy D. 2005. “Who Toes the Party Line? Cues, Values, and Individual Differences.” *Political Behavior* 27 (2). Springer: 163–82.

Kinder, Donald R, and David O Sears. 1981. “Prejudice and Politics: Symbolic Racism Versus Racial Threats to the Good Life.” *Journal of Personality and Social Psychology* 40 (3). American Psychological Association: 414.

Kramon, Eric, and Keith Weghorst. 2012. “List Experiments in the Field.” In *Newsletter of the Apsa Experimental Section*.

———. 2019. “(Mis) Measuring Sensitive Attitudes with the List Experiment: Solutions to List Experiment Breakdown in Kenya.” *Public Opinion Quarterly* 83 (S1). Oxford University Press UK: 236–63.

Krosnick, Jon A. 1991. “Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys.” *Applied Cognitive Psychology* 5 (3). Wiley Online Library: 213–36.

Kuklinski, James H, Paul M Sniderman, Kathleen Knight, Thomas Piazza, Philip E Tetlock, Gordon R Lawrence, and Barbara Mellers. 1997. “Racial Prejudice and Attitudes Toward Affirmative Action.” *American Journal of Political Science*. JSTOR, 402–19.

Lavrakas, Paul J, Michael W Traugott, Courtney Kennedy, Allyson L Holbrook, Edith D de Leeuw, and Brady T West. 2019. *Experimental Methods in Survey Research: Techniques That Combine Random Sampling with Random Assignment*. John Wiley & Sons.

Leary, Mark R, and Robin M Kowalski. 1990. “Impression Management: A Literature Review and Two-Component Model.” *Psychological Bulletin* 107 (1). American Psychological Association: 34.

Leeper, Thomas J, Sara B Hobolt, and James Tilley. 2020. “Measuring Subgroup Preferences in Conjoint Experiments.” *Political Analysis* 28 (2). Cambridge University Press: 207–21.

Lensvelt-Mulders, Gerty JLM, Joop J Hox, Peter GM Van der Heijden, and Cora JM Maas. 2005. “Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation.” *Sociological Methods & Research* 33 (3). Sage Publications Sage CA: Thousand Oaks, CA: 319–48.

Lyall, Jason, Graeme Blair, and Kosuke Imai. 2013. “Explaining Support for Combatants During Wartime: A Survey Experiment in Afghanistan.” *American Political Science Review* 107 (4). Cambridge University Press: 679–705.

Macrae, C Neil, Galen V Bodenhausen, Alan B Milne, and Jolanda Jetten. 1994. “Out of Mind but Back in Sight: Stereotypes on the Rebound.” *Journal of Personality and Social Psychology* 67 (5). American Psychological Association: 808.

Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton University Press.

Nicholson, Stephen P. 2011. “Dominating Cues and the Limits of Elite Influence.” *The Journal of Politics* 73 (4). Cambridge University Press New York, USA: 1165–77.

———. 2012. “Polarizing Cues.” *American Journal of Political Science* 56 (1). Wiley Online Library: 52–66.

Nisbett, Richard E, and Timothy D Wilson. 1977. “Telling More Than We Can Know: Verbal Reports on Mental Processes.” *Psychological Review* 84 (3). American Psychological Association: 231.

Norenzayan, Ara, and Norbert Schwarz. 1999. “Telling What They Want to Know: Participants Tailor Causal Attributions to Researchers’ Interests.” *European Journal of Social Psychology* 29 (8). Wiley Online Library: 1011–20.

Nosek, Brian A. 2007. “Implicit–Explicit Relations.” *Current Directions in Psychological Science* 16 (2). SAGE Publications Sage CA: Los Angeles, CA: 65–69.

Rankin, Robert E, and Donald T Campbell. 1955. “Galvanic Skin Response to Negro and White Experimenters.” *The Journal of Abnormal and Social Psychology* 51 (1). American Psychological Association: 30.

Riambau, Guillem, and Kai Ostwald. 2019. “Placebo Statements in List Experiments: Evidence from a Face-to-Face Survey in Singapore.” *Political Science Research and Methods*. Cambridge University Press, 1–8.

Rosenfeld, Bryn, Kosuke Imai, and Jacob N Shapiro. 2016. “An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions.” *American Journal of Political Science* 60 (3). Wiley Online Library: 783–802.

Schwarz, Norbert. 1999. “Self-Reports: How the Questions Shape the Answers.” *American Psychologist* 54 (2). American Psychological Association: 93.

Schwarz, Norbert, and Gerald L Clore. 1983. “Mood, Misattribution, and Judgments of Well-Being: Informative and Directive Functions of Affective States.” *Journal of Personality and Social Psychology* 45 (3). American Psychological Association: 513.

Simon, Herbert, and James March. 2006. “Administrative Behavior and Organizations.” *Organizational Behavior 2: Essential Theories of Process and Structure* 2 (41). ME Sharpe.

Sniderman, Paul M. 2018. “Some Advances in the Design of Survey Experiments.” *Annual Review of Political Science* 21. Annual Reviews: 259–75.

Sniderman, Paul M, Thomas Piazza, Philip E Tetlock, and Ann Kendrick. 1991. “The New Racism.” *American Journal of Political Science*. JSTOR, 423–47.

Steiner, Peter M, Christiane Atzmüller, and Dan Su. 2016. “Designing Valid and Reliable Vignette Experiments for Survey Research: A Case Study on the Fair Gender Income Gap.” *Journal of Methods and Measurement in the Social Sciences* 7 (2): 52–94.

Warner, Stanley L. 1965. “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias.” *Journal of the American Statistical Association* 60 (309). Taylor & Francis: 63–69.

Winters, Matthew S, and Rebecca Weitz-Shapiro. 2013. “Lacking Information or Condoning Corruption: When Do Voters Support Corrupt Politicians?” *Comparative Politics* 45 (4). City University of New York: 418–36.

Yu, Jun-Wu, Guo-Liang Tian, and Man-Lai Tang. 2008. “Two New Models for Survey Sampling with Sensitive Characteristic: Design and Analysis.” *Metrika* 67 (3). Springer: 251.

Zigerell, Lawrence J. 2011. “You Wouldn’t Like Me When I’m Angry: List Experiment Misreporting.” *Social Science Quarterly* 92 (2). Wiley Online Library: 552–62.