JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 00, NO. 0, 1–15: Applications and Case Studies
https://doi.org/10.1080/01621459.2019.1585248

Design, Identification, and Sensitivity Analysis for Patient Preference Trials

Dean Knox[a], Teppei Yamamoto[b], Matthew A. Baum[c], and Adam J. Berinsky[b]

[a]Department of Politics, Princeton University, Princeton, NJ; [b]Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA; [c]John F. Kennedy School of Government, Harvard University, Cambridge, MA

ABSTRACT

Social and medical scientists are often concerned that the external validity of experimental results may be compromised because of heterogeneous treatment effects. If a treatment has different effects on those who would choose to take it and those who would not, the average treatment effect estimated in a standard randomized controlled trial (RCT) may give a misleading picture of its impact outside of the study sample. Patient preference trials (PPTs), where participants' preferences over treatment options are incorporated in the study design, provide a possible solution. In this paper, we provide a systematic analysis of PPTs based on the potential outcomes framework of causal inference. We propose a general design for PPTs with multi-valued treatments, where participants state their preferred treatments and are then randomized into either a standard RCT or a self-selection condition. We derive nonparametric sharp bounds on the average causal effects among each choice-based subpopulation of participants under the proposed design. We also propose a sensitivity analysis for the violation of the key ignorability assumption sufficient for identifying the target causal quantity. The proposed design and methodology are illustrated with an original study of partisan news media and its behavioral impact. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

ARTICLE HISTORY
Received March 2017
Accepted January 2019

KEYWORDS
Causal inference; External validity; Nonparametric bounds; Randomized controlled trial.

1. Introduction

Randomized controlled trials (RCTs) are widely used in the social and medical sciences to estimate the causal effects of treatments of interest. The random assignment of treatments ensures the internal validity of the study, in the sense that observed differences in the distribution of outcomes between randomized treatment groups can be interpreted as causal effects of the treatments. Carefully controlled randomization, however, often comes at the cost of external validity. That is, conclusions from RCTs may not generalize to situations outside of that particular experiment. Without sufficient external validity, RCTs are not informative about the substantive, real-world questions in which scientists and practitioners are ultimately interested.

In RCTs, preferences of experimental subjects over treatment options often play an important role. Even in a well-controlled study on a representative sample from the target population, heterogeneity of treatment effects across treatment preferences may limit the study's externally validity. For example, a medical treatment that was found to be ineffective on average in a RCT may in fact be highly beneficial for the patients who would choose to take it if they were able to. In a standard RCT, however, researchers cannot make such nuanced inferences because all subjects are forced to take treatments randomly chosen by the researchers.

In this article, we propose a new experimental design for patient preference trials (PPTs), in which subjects' preferences over treatments are systematically incorporated in the study design. The proposed design consists of two stages of randomization and synthesizes many of the variants of PPTs previously used in social (Gaines and Kuklinski 2011; Arceneaux, Johnson, and Murphy 2012) and medical (King et al. 2005; Howard and Thornicroft 2006) applications. First, all participants state their preferred treatments prior to entering the study. Then, we randomize them into either a standard RCT or a self-selection condition. In the latter condition, they choose the treatment as they would in the real world. Finally, we measure the outcome variables of interest. The proposed design is novel in that it allows the researcher to incorporate in the analysis the discrepancy between subjects' stated preferences and their actual choice of treatments. This modification is important because respondents to a survey question often fail to report their underlying preferences to the interviewer, whether consciously or subconsciously.

Using the potential outcomes framework of causal inference (Neyman 1923; Rubin 1974), we define a causal quantity which we call the average choice-specific treatment effect (ACTE), representing the conditional average treatment effect for the subpopulation of subjects who would choose a particular treatment option. We show that the point identification of this quantity for

CONTACT Teppei Yamamoto *teppei@mit.edu* Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

These materials were reviewed for reproducibility.

a multi-valued treatment requires the strong assumption that the discrepancy between stated preference and actual choice is ignorable. That is, the naïve estimate of the ACTE using stated preference will be biased if the measurement errors are systematically correlated with unobserved characteristics that affect the potential outcomes.

To make valid inference about the ACTE under the proposed design, we derive nonparametric sharp bounds on this causal quantity without assuming the ignorability of the measurement error in stated treatment preferences. We also propose a sensitivity analysis where we quantify the assumed informativeness of the stated preferences about actual choices via a sensitivity parameter and analyze how the ACTE responds to the change in this parameter. Finally, we develop a simulation-based procedure to make statistical inference for the bounds and sensitivity results in small samples. We provide open-source software, `ppt`, for implementing the proposed methodology.

We illustrate the proposed design and methodology with an original survey experiment, where we investigate the effect of partisan political news media on the subjects' perception of the media and their subsequent political behavior. Our primary interest is in how the effects of our treatments vary depending on whether subjects would actually consume such partisan media if they could choose to do so.

Despite the prevalence of PPTs across scientific disciplines, very few methodological investigations have been conducted on the topic from the perspective of causal inference. A notable exception is Long, Little, and Lin (2008), who employed a framework similar to ours to define causal quantities. For the identification and estimation of those quantities, however, they assume a parametric model between the unobserved choice and observed covariates for participants in the self-selection condition and use an EM algorithm to estimate the causal quantities as functions of model parameters. In contrast, our model-free approach avoids distributional or functional-form assumptions for better credibility of the resulting inference.

The rest of the article proceeds as follows. Section 2 describes the background motivation of the empirical example. Section 3 formally describes the proposed design and defines causal quantities of interest and assumptions. Sections 4–6 discuss the proposed methodology. Section 7 applies the method to the empirical example. Section 8 investigates finite-sample performance of the proposed inferential approach via Monte Carlo simulations. Section 9 concludes.

2. A Motivating Example

In recent years, many scholars (e.g., Prior 2007) have explored the political consequences of increased media choice in the 21st century. The explosion of media outlets has vastly increased the choices available to consumers and allowed for the development of ideological "niche" news programming (Hamilton 2005). A great deal of research has sought to determine the effects of this unprecedented media fragmentation (e.g., Stroud 2011; Kim 2009; Iyengar and Hahn 2009; Levendusky 2013).

Among several significant strands of this research program, a predominant body of scholarship has sought to delineate the effects of consuming ideologically polarized media on attitudes toward the broader mass media. According to Gallup (2014), between 1976 and 2014, the percentage of Americans expressing "a great deal" or "a fair amount" of trust in the media fell from 72% to 44%. People who distrust the media may conclude it cannot report in an unbiased manner. As a result, the public may dismiss media content as unreliable and increasingly become suspicious of and antagonistic toward the news media more generally (Arceneaux, Johnson, and Murphy 2012; Ladd 2012).

To explore this phenomenon, we conducted an experiment in June 2014 on a sample of 3023 American adults, recruited by Survey Sampling International (SSI). Our goal was to estimate the effect of exposing subjects to pro- and counter-attitudinal political news programming (as opposed to nonpolitical entertainment shows) on their sentiment toward specific news programs and the media in general. We also explored whether such programming produces behavioral responses, such as changes in propensity to discuss the content of the media with friends. Specifically, we selected a short clip from each of the following television programs: (1) The Rachel Maddow Show (MSNBC), (2) Jamie's Kitchen with Jamie Oliver (Food Network), (3) Dirty Jobs with Mike Rowe (Discovery Channel), and (4) The O'Reilly Factor with Bill O'Reilly (Fox News). We carefully selected clips from the two political shows—Rachel Maddow and The O'Reilly Factor—to match as closely to each other in topic and content as possible. We selected clips that focused on energy policy (specifically, the Obama administration's policies regarding domestic energy production and their effects on gas prices). Finally, we merged the two entertainment shows into a single treatment condition ("entertainment") in our analysis.

One of our primary concerns in the design of our study was that the existing experimental studies of partisan media effects had limited external validity because they paid inadequate attention to the preferences of subjects over treatment options. Namely, the average treatment effect obtained in a standard RCT could mask fundamental heterogeneity across different types of individuals and misrepresent the overall impact of media polarization in the "real" political world. For instance, it could be the case that partisan news is highly persuasive for some people—say, those least likely to consume it in the real world—while having little or no persuasive effect among people who are most likely to consume it.

A natural approach to incorporating preferences is to adopt one of the commonly used PPT designs. For example, Arceneaux, Johnson, and Murphy (2012) conducted a similar media choice experiment in which respondents were asked their news preferences before being randomly assigned to a particular treatment condition. A PPT based on the measurement of stated preferences like this, however, appears inadequate in our context. This is because research has shown that people often have difficulty assessing what they would actually do or prefer (Clausen 1968) or have done in the past (Prior 2009) when offered a hypothetical choice or asked about past behavior. Theories regarding the source of this gap between self-reported preferences and actual behavior, like media consumption, are manifold. These theories range from a bias toward offering socially desirable responses on topics like voting (Rogers and Aida 2013) and sensitive topics (Brown and Sinclair 1999; Hser, Maglione, and Boyle 1999; Payne 2010); to selective retention of pro-attitudinal information (Campbell et al. 1960) or motivated

reasoning (Levendusky 2013); to an inability to accurately remember prior behavior (Tourangeau 1999).

Given these considerations about the inadequacy of existing experimental designs, we implemented a new PPT design which we will describe in Section 3. In Section 7, we present results from our analysis of this experiment with our proposed methodology.

3. Design and Assumptions

In this section, we introduce the notation required for our methodology. We define our causal quantities of interest and discuss their substantive interpretations. We then introduce several assumptions for identification analysis.

3.1. Notation and the Proposed Design

Suppose that we have a random sample of N experimental subjects from the population of interest. We consider a study where the goal is to estimate the effect of a J-valued treatment on an outcome of interest. Let $A_i \in \mathcal{A} \equiv \{0, 1, \ldots, J-1\}$ denote the treatment that subject i actually receives in the study. For the rest of the paper, we call this the "actual treatment," or simply the "treatment" when the meaning is obvious from the context. Without loss of generality, we impose the standard total ordering on $\mathcal{A}$.

Our proposed design for PPTs proceeds as follows. First, all N subjects in the study sample are asked to state their preferred treatment, $S_i \in \mathcal{A}$. Second, after an optional "washout" period, or a set of additional questions as we discuss in Section 7.1, the subjects are randomized into one of the two conditions: either they will be forced to take the randomly assigned treatment, or they will be allowed to freely choose the treatment of their own accord. We refer to this actual choice as $C_i \in \mathcal{A}$. Formally, we use the "design indicator" $D_i \in \{0, 1\}$ to denote whether subject i is in the forced-exposure condition ($D_i = 1$) or the free-choice condition ($D_i = 0$). Third, the subjects then receive treatment (A_i) according to the protocol determined by their design indicator. That is, A_i is randomized if and only if $D_i = 1$; in this case, C_i is unobserved. For the subjects with $D_i = 0$, their treatments equal the treatments they have chosen. Therefore, we have $A_i = C_i$ if $D_i = 0$. Finally, the outcome of interest, $Y_i \in \mathcal{Y}$, is measured for every subject.

Under the proposed design, the potential outcome for subject i can be defined as $Y_i(a) \in \mathcal{Y}$. This represents the value of the outcome of interest that would be realized if i received the treatment $a \in \mathcal{A}$. By this notation, we are implicitly making the stable unit treatment value assumption (SUTVA; Rubin 1990), which posits that subjects cannot be affected by the treatments received by any other subjects (no interference) and that subjects exhibit the same value of the outcome no matter how the treatment $A_i = a$ was received (stability or consistency). In particular, the notation assumes that there is no design effect, that is, the potential outcomes remain stable across the two design conditions. Long, Little, and Lin (2008) called this assumption the exclusion restriction. The no-design-effect assumption would be violated if, for example, a nominally identical treatment had different effects on the outcome for the same unit depending on whether the treatment was randomly assigned in the forced-exposure condition or voluntarily chosen in the free-choice condition. Under the SUTVA, we can express the observed outcome as $Y_i = \sum_{a \in \mathcal{A}} Y_i(a)\mathbf{1}\{A_i = a\} = Y_i(A_i)$ for any i. The cdf of $Y_i(a)$ is denoted by $F_{Y(a)}(y) = \Pr(Y_i(a) \leq y)$.

The diagram in Figure 1 graphically summarizes the proposed design. Several important features of this design are worth mentioning. First, the proposed design combines the standard RCT ($D_i = 1$, upper arm) with a pure self-selection study ($D_i = 0$, lower arm) via randomization. As discussed in Section 4, this allows us to infer more about the unobserved choice behavior of the subjects who are assigned to the forced-exposure condition. Second, our design clearly distinguishes the stated preference of the subjects (S_i) from their actual choice (or "revealed preferences," as they are often called in the social sciences, C_i). As pointed out in Section 2, social and medical scientists are often concerned that stated preferences may be unreliable due to various sources of systematic measurement error. Thus, a "naïve" analysis that takes the stated preferences at their face value and ignores the possible measurement error may lead to an invalid estimate. Finally, note that we allow the treatment variable to be multi-valued, instead of binary. In fact, as previously shown by Long, Little, and Lin (2008) and Gaines and Kuklinski (2011) and revisited in Section 4, assuming a binary treatment greatly simplifies the problem, leading to point identification of the ACTEs (defined shortly). However, as in the media choice example, social and medical scientists are often interested in testing the effects of more than two treatments in a single study.

There exist many previous studies in both social and medical sciences that use PPT designs related to ours (King et al. 2005; Kowalski and Mrdjenovich 2013). For example, a commonly used design in clinical trials (Brewin and Bradley 1989) begins by asking subjects whether they have a preference between treatment options and then randomizing only those subjects who expressed no preference. A close variant of this design (Schmoor, Olschewski, and Schumacher 1996) asks participants whether they agree to be randomized into treatment conditions. These designs may be preferable in situations where randomizing subjects against their preferences is practically or ethically infeasible. However, these designs are critically limited in terms of external validity because the treatment effects are only identified for subjects who have no (or weak) treatment preferences.

The key advantage of our proposed design is the combination of stated preference measurement for all subjects *and* randomization into the forced-exposure and free-choice conditions, which has never been implemented as far as we are aware. For example, a popular approach that involves measurement of stated preferences at the beginning of study forces all respondents to randomly assigned treatments (Torgerson, Klaber-Moffett, and Russell 1996), whereas studies that contain randomization into forced-exposure and free-choice arms typically omit stated preference measurement (Gaines and Kuklinski 2011). The importance of measuring stated preferences before randomization is worth emphasizing, for (as formally discussed in Section 4) the variable enables more informative inference across the forced-exposure and free-choice arms. For

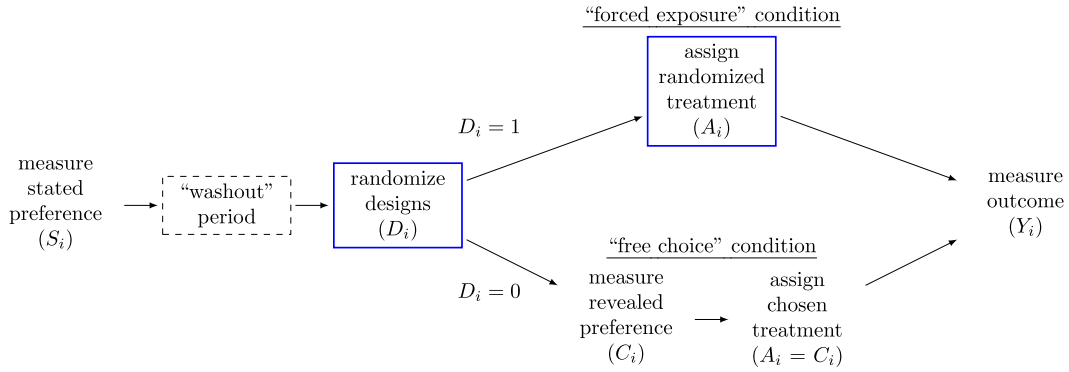

Figure 1. Diagram of the proposed PPT design. In the proposed design, subjects are first asked to state preferences about the treatment options (S_i) and (after an optional "washout" period) randomized into design conditions (D_i). In the "forced exposure" arm (top, $D_i = 1$), subjects are randomly assigned to treatments irrespective of their stated preferences (A_i). In the "free choice" arm (bottom, $D_i = 0$), the subjects are asked to choose the treatment they want to take (C_i) and actually exposed to that treatment ($A_i = C_i$). Finally, the outcome measure is taken on all subjects (Y_i). In the diagram, the blue boxes indicate random assignment and the dashed box indicates an optional component.

example, Arceneaux, Johnson, and Murphy (2012) report results from a series of RCTs, one of which included measurement of stated preferences and another which involved randomization into a free-choice condition. Unfortunately, these two studies are conducted separately on populations with possibly different characteristics, rendering inference from combined data difficult.

3.2. Quantities of Interest

A common causal quantity of interest in the social and medical sciences is the (population) *average treatment effect* (*ATE*), which is defined as follows.

$$\delta(a, a') \equiv \mathbb{E}[Y_i(a) - Y_i(a')],$$

for any a and $a' \in \mathcal{A}$. This quantity represents the (additive) causal effect of treating a unit with treatment a as opposed to treatment a', averaged unconditionally over the sampling distribution. It is widely known that the ATE can be nonparametrically identified in a standard RCT, where both treatments a and a' are randomly assigned with nonzero probabilities, and can be estimated with very simple estimators such as the difference-in-means.

However, the ATE is often not the only causal parameter that is of substantive interest in a given applied setting. For example, in the media choice experiment introduced in Section 2, our interest was not only in the average effect of exposing every American adult to one program versus another, but also in investigating heterogeneity in media effects based on the respondents' likely media consumption in the real world. Likewise, in a medical application, researchers may want to study whether a new treatment has beneficial effects on the patients who would actually choose to use the treatment, or whether it may have a potential harmful impact on patients if it is applied in spite of a diverging preference.

In the rest of this article, we focus on an alternative causal quantity which addresses these more nuanced questions,

$$\tau(a, a'|c) \equiv \mathbb{E}[Y_i(a) - Y_i(a')|C_i = c], \quad (1)$$

for any a, a', and $c \in \mathcal{A}$. We call this quantity the *ACTE*. The ACTE represents the average effect of treating a unit with treatment a instead of a' among the units who would choose treatment c if they were allowed to. For example, in the media choice experiment, we may be interested in the effect of watching a pro-attitudinal news program (a) instead of an entertainment show (a') among those who would actually be watching entertainment when they were freely choosing the programs to watch ($c = a'$). Similarly, a psychiatrist may want to estimate the potentially adverse effect of imposing a new therapy on patients who would prefer to keep to the old treatment. Thus, the ACTE is useful for the investigation of substantively meaningful heterogeneity in treatment effects in a "natural" condition, where units would be choosing treatments without an intervention from researchers. Note that, as expected, the overall ATE can be expressed as the weighted average of the ACTEs, where the weights are given by the proportions of units who would choose each of the treatment options (i.e., $\delta(a, a') = \sum_c \tau(a, a'|c) \Pr(C_i = c)$).

The ACTE has a close connection with the more commonly used *ATE on the treated (ATT)*, defined as follows.

$$\gamma(a, a') \equiv \mathbb{E}[Y_i(a) - Y_i(a')|A_i = a],$$

for a and $a' \in \mathcal{A}$. The ATT represents the average effect of treatment a versus a' among those units who are actually treated with a. Conventionally in the literature, *how* those units come to be actually treated with a is left implicit in the definition of this quantity. For example, in a standard RCT where treatments are randomly assigned and imposed, the ATT is equivalent to the ATE because A_i is statistically independent of the potential outcomes (i.e., $\gamma(a, a') = \delta(a, a')$ for any $a, a' \in \mathcal{A}$). On the other hand, in the so-called encouragement design where an encouragement (or "instrument") for taking a particular treatment option is randomized (e.g., Hirano et al. 2000), the actual treatment status A_i reflects the subject's voluntary action of choosing to take the treatment and the ATT now has a substantive meaning similar to the ACTE. This implies that the substantive interpretation of the ATT as a causal quantity crucially depends on the study design. In this article, we opt to introduce the new causal quantity ACTE because its interpretation is clearer and less affected by auxiliary design assumptions than the ATT.

3.3. Assumptions

Here, we introduce a set of statistical assumptions and discuss their relationships with the design we propose. Note that the proposed design involves two random assignments. First, the randomization of subjects into the forced-exposure and free-choice conditions implies the following assumption.

Assumption 1 (*Randomization of designs*).

$$\{Y_i(a),\ C_i,\ S_i\} \perp\!\!\!\perp D_i \quad \text{for all} \quad a \in \mathcal{A}.$$

Long, Little, and Lin (2008) referred to this assumption as "no selection bias from randomization." Second, in the forced-exposure condition, the treatments are randomly assigned and imposed on each subject. This implies that the following assumption is also guaranteed to be true.

Assumption 2 (*Randomization of the forced treatment*).

$$\{Y_i(a),\ C_i,\ S_i\} \perp\!\!\!\perp A_i \mid D_i = 1 \quad \text{for all} \quad a \in \mathcal{A}.$$

In addition to these design-guaranteed assumptions, existing studies using PPTs often make the following untestable assumption (e.g., Arceneaux, Johnson, and Murphy 2012).

Assumption 3 (*Mean ignorability of measurement error*).

$$\mathbb{E}[Y_i(a)|C_i = c] = \mathbb{E}[Y_i(a)|S_i = c] \quad \text{for any} \quad a, c \in \mathcal{A}.$$

This assumption states that the potential outcomes of the units who would choose a particular treatment option are on average equal to the potential outcomes of the (potentially different) set of units who state that they would choose the same treatment. In other words, Assumption 3 holds if the discrepancy between the stated and revealed preferences (which one may call the measurement error if the stated preference is thought of as a measure of underlying preference) is ignorable. The assumption may be violated if the discrepancy between the stated preference and actual choice is systematically correlated with any background characteristic of the units that are associated with the potential outcomes.

Assumption 3 is not directly testable because the conditional expectation on the left-hand side is unobservable for $a \neq c$. However, Assumption 3 has two empirical implications which can be tested with observed information. First, Assumptions 1–3 jointly imply the following relationship.

$$\mathbb{E}[Y_i|A_i = a, D_i = 0] = \mathbb{E}[Y_i|A_i = S_i = a, D_i = 1], \tag{2}$$

for any $a \in \mathcal{A}$. Second, for outcomes that are bounded from below ($\underline{y}$) and above ($\overline{y}$), it can be shown that the following inequalities must hold under Assumptions 1–3.

$$\begin{aligned} \underline{y} \leq\ & \frac{\mathbb{E}[Y_i|A_i = a, D_i = 0]}{-\mathbb{E}[Y_i|A_i = S_i = a, D_i = 0]\Pr(A_i = a|S_i = a, D_i = 0)} \\ & \leq \overline{y} \end{aligned} \tag{3}$$

for any $a \in \mathcal{A}$. Proofs are provided in Web Appendix A1.

Assumption 3 may be attractive because, as we show in Section 4, it allows the point identification of the ACTE only with the forced-exposure condition. By making Assumption 3, the researcher can save the cost of employing an additional experimental arm. However, the assumption is a strong one in many applied contexts, as we discussed in Sections 2 and 3.1. In such applications, we recommend against dropping the free-choice condition entirely, and also recommend that the above observable implications of the assumption be tested with the collected data before the assumption is made in the analysis. Tests can be conducted in the usual manner based on the sample analogues of the expressions and their asymptotic sampling properties, obtained via standard techniques like the delta method.

4. Nonparametric Identification Analysis

In this section, we present the results of our nonparametric identification analysis for the ACTE. Our goal is to find nonparametric bounds on $\tau(a, a'|c)$ as functions of the joint distribution of $\{Y_i, D_i, A_i, S_i\}$, which are always observed for all $i \in \{1, \ldots, N\}$ under the proposed design. First, we consider the identifiability of the ACTE when we only make the assumptions that are guaranteed to hold by the study design (i.e., Assumptions 1 and 2) as well as the SUTVA and no design effect. In Web Appendix A2, we show that the ACTE can be expressed as follows under those assumptions.

$$\begin{aligned} &\tau(a, a'|c) \\ &= \frac{1}{\Pr(A_i = c|D_i = 0)} \\ &\quad \times \left\{ \begin{array}{l} \mathbb{E}[Y_i|A_i = a, D_i = 1] - \mathbb{E}[Y_i|A_i = a', D_i = 1] \\ -\mathbb{E}[Y_i|A_i = a, D_i = 0]\Pr(A_i = a|D_i = 0) \\ +\mathbb{E}[Y_i|A_i = a', D_i = 0]\Pr(A_i = a'|D_i = 0) \\ -\sum_{c' \notin \{a,c\}} \mathbb{E}[Y_i(a)|C_i = c']\Pr(A_i = c'|D_i = 0) \\ +\sum_{c' \notin \{a',c\}} \mathbb{E}[Y_i(a')|C_i = c']\Pr(A_i = c'|D_i = 0) \end{array} \right\}, \end{aligned} \tag{4}$$

for any a, a', and $c \in \mathcal{A}$. Equation (4) immediately gives us three important results. First, Equation (4) contains as many as $2(J - 2)$ terms (when $a \neq a' \neq c$) that cannot be identified from observed data under Assumptions 1 and 2. When $a \neq a' = c$ or $a = c \neq a'$, some of these can replaced with observed quantities, but $J - 2$ unidentified terms remain. Thus, it can be concluded that the ACTE is unidentified by the proposed PPT design itself.

Second, when the treatment is binary as in many social and medical RCTs (i.e., $J = 2$), the unidentified terms drop out of Equation (4). This implies that the ACTE is point-identified under Assumptions 1 and 2 alone if $J = 2$, and is written as

$$\tau(a, a'|c) = \left\{ \begin{array}{ll} \frac{\mathbb{E}[Y_i|D_i=0] - \mathbb{E}[Y_i|A_i=a', D_i=1]}{\Pr(A_i=a|D_i=0)} & \text{if} \quad c = a, \\ \frac{\mathbb{E}[Y_i|A_i=a, D_i=1] - \mathbb{E}[Y_i|D_i=0]}{\Pr(A_i=a'|D_i=0)} & \text{if} \quad c = a', \end{array} \right.$$

for a, a', and $c \in \{0, 1\}$. This exactly matches Gaines and Kuklinski's (2011, p. 729) result, where they consider a PPT design that is identical to ours except that it does not contain the measurement of stated preferences S_i and only the $J = 2$ case is considered. The same result is also obtained by Long, Little, and Lin (2008) using a framework more similar to ours. Thus,

we verify their earlier result under the current framework and also show that our proposed framework encompasses theirs as a special case.

Third, if we make Assumption 3 in addition to Assumptions 1 and 2, the unidentified terms in Equation (4) become identified as $\mathbb{E}[Y_i(a'')|C_i = c'] = \mathbb{E}[Y_i|A_i = a'', S_i = c', D_i = 1]$ for $a'' \in \{a, a'\}$. This implies that the ACTE can be point identified for any J under Assumptions 1–3 and is given by the following expression.

$$\begin{aligned}\tau(a, a'|c) = \mathbb{E}[Y_i|S_i = c, A_i = a, D_i = 1] \\ -\mathbb{E}[Y_i|S_i = c, A_i = a', D_i = 1],\end{aligned} \tag{5}$$

for a, a', and $c \in \mathcal{A}$. Equation (5) makes it clear that the forced-exposure group alone is sufficient for the identification of the ACTE when we make Assumptions 2 and 3. Indeed, Arceneaux, Johnson, and Murphy (2012, pp. 182–183) use Equation (5) to estimate the ACTE in their experiment, which consisted of the forced-exposure arm of our proposed design alone. As we discussed in Section 3, while this design choice may be reasonable in some applied contexts, it must be made with caution because Assumption 3 is strong and omitting the free-choice condition precludes the testing of its observable implications. From here on, we call Equation (5) the "naïve estimator" of the ACTE.

What if we want to avoid Assumption 3 or analyze nonbinary treatments? In this case, the forced-exposure arm alone is completely uninformative about the ACTE. We present two partial identification results using both forced-exposure and free-choice arms, which provide *sharp bounds* (i.e., the tightest possible given all the observed information; Manski 1995) on $\tau(a, a'|c)$ under various scenarios.

4.1. General Results

Our first set of results, summarized in Proposition 1, is the more general of the two and valid for any real-valued outcome ($\mathcal{Y} \subseteq \mathbb{R}$) under the proposed design.

Proposition 1 (Nonparametric sharp bounds on the ACTE). Under Assumptions 1 and 2, $\tau(a, a'|c)$ can be partially identified at least up to the following nonparametric bounds

$$\begin{aligned}&\sum_{s\in\mathcal{A}}\left(\lim_{y^*\to-\infty}\right. \\ &\left[\int_{y^*}^{\infty}\max\left\{0, 1-\frac{\begin{array}{c}1-F(y|s,a',1)\\ +\left\{1-F(y|s,a',0)\right\}P(a'|s,0)\end{array}}{\Pr(A_i=c|S_i=s,D_i=0)}\right\}\right. \\ &\left.\left.-\min\left\{1, \frac{F(y|s,a,1)-F(y|s,a,0)P(a|s,0)}{\Pr(A_i=c|S_i=s,D_i=0)}\right\}\,\mathrm{d}y\right]\right) \\ &\times\Pr(S_i=s|A_i=c,D_i=0) \\ \le\ &\tau(a,a'|c)\ \le \\ &\sum_{s\in\mathcal{A}}\left(\lim_{y^*\to-\infty}\right. \\ &\left[\int_{y^*}^{\infty}\min\left\{1, \frac{F(y|s,a',1)-F(y|s,a',0)P(a'|s,0)}{\Pr(A_i=c|S_i=s,D_i=0)}\right\}\right. \\ &\left.\left.-\max\left\{0, 1-\frac{\begin{array}{c}1-F(y|s,a,1)\\ +\left\{1-F(y|s,a,0)\right\}P(a|s,0)\end{array}}{\Pr(A_i=c|S_i=s,D_i=0)}\right\}\,\mathrm{d}y\right]\right) \\ &\times\Pr(S_i=s|A_i=c,D_i=0),\end{aligned} \tag{6}$$

where $F(y|s,a,d) = \Pr(Y_i \le y|S_i = s, A_i = a, D_i = d)$ and $P(a|s,0) = \Pr(A_i = a \mid S_i = s, D_i = 0)$ for any a, a', and $c \in \mathcal{A}$. If $a' = c$, these bounds are sharp and simplify to the following expression

$$\begin{aligned}&\sum_{s\in\mathcal{A}}\left(\lim_{y^*\to-\infty}\left[\int_{y^*}^{\infty}1\right.\right. \\ &\left.\left.-\min\left\{1, \frac{\begin{array}{c}F(y|s,a,1)\\ -F(y|s,a,0)P(a|s,0)\end{array}}{\Pr(A_i=c|S_i=s,D_i=0)}\right\}\,\mathrm{d}y+y^*\right]\right) \\ &\times\Pr(S_i=s|A_i=c,D_i=0) \\ &-\mathbb{E}[Y_i|A_i=c,D_i=0] \\ \le\ &\tau(a,c|c)\ \le \\ &\sum_{s\in\mathcal{A}}\left(\lim_{y^*\to-\infty}\left[\int_{y^*}^{\infty}1\right.\right. \\ &\left.\left.-\max\left\{0, 1-\frac{\begin{array}{c}1-F(y|s,a,1)\\ +\left\{1-F(y|s,a,0)\right\}P(a|s,0)\end{array}}{\Pr(A_i=c|S_i=s,D_i=0)}\right\}\,\mathrm{d}y\right.\right. \\ &\left.\left.+y^*\right]\right)\cdot\Pr(S_i=s|A_i=c,D_i=0) \\ &-\mathbb{E}[Y_i|A_i=c,D_i=0],\end{aligned} \tag{7}$$

for any a and $c \in \mathcal{A}$.

Intuitively, the observed forced-choice distribution of $Y_i(a)$ is a mixture of the choice-specific component distributions, where size of each component is known but the choice-specific distribution is unobserved for all $C_i \neq A_i$. We bound these unobserved component distributions by first applying the Fréchet–Hoeffding bounds on the joint distribution of $Y_i(a)$ and C_i for each observed stratum defined by S_i, conditional on $C_i \neq a$.

Then, sharp bounds on $\mathbb{E}[Y_i(a) \mid C_i = c]$ are derived. A detailed proof can be found in Web Appendix A3.

We offer several remarks on Proposition 1. First, the bounds on $\tau(a, a'|c)$ are tighter when more units choose the treatment of interest (c) in the free-choice condition. This is because, intuitively, the worst-case assumptions for the unobserved potential outcomes apply to a smaller portion of the population. Second, we can prove the bounds to be sharp only when $a' = c$, that is, when one of the average potential outcomes in $\tau(a, a'|c)$ can be point identified from the observed outcome for the free-choice group. This limitation motivates our second set of identification results.

4.2. Sharp Bounds for Binary Outcomes

Next, we restrict analysis to outcome variables that are binary ($\mathcal{Y} \in \{0, 1\}$) and derive another set of nonparametric bounds on the ACTE. In this case, we can obtain the sharp bounds on $\tau(a, a'|c)$ for any a, a' and $c \in \mathcal{A}$ (in particular, even when $a \neq a' \neq c$) by incorporating the full joint distribution of the observed variables in the derivation of the bounds. This is achieved via the linear programming approach based on principal stratification (Balke 1995; Balke and Pearl 1997; Frangakis and Rubin 2002), which has recently been used for nonparametric identification analysis of various causal quantities (e.g., Yamamoto 2012; Imai, Tingley, and Yamamoto 2013). First, we define $2^J J^2$ principal strata, a partition of the population of units based on the values of their potential outcomes $(Y_i(0), \ldots, Y_i(J-1))$ as well as the values of their stated and revealed preferences (S_i and C_i). Then we consider the population proportion of each principal stratum, which we denote by $\phi_{y_0,\ldots,y_{J-1},s,c} \equiv \Pr(Y_i(0) = y_0, \ldots, Y_i(J-1) = y_{J-1}, S_i = s, C_i = c)$, where $y_0, \ldots, y_{J-1} \in \{0, 1\}$ and $s, c \in \mathcal{A}$. For the rest of this section, we focus on the case of a tri-valued treatment ($J = 3$, as in the media choice example) for notational tractability, although the proposed method can be applied more generally. There are a total of 72 unique principal strata when $J = 3$, corresponding to unique combinations in the indices of $\phi_{y_0,y_1,y_2,s,c}$. The proposed method can also be applied to nonbinary categorical outcomes with a straightforward extension, which we do not pursue here to keep the exposition simple.

The following proposition shows that the sharp bounds on the ACTE can be obtained by solving a linear programming problem when the outcome is binary.

Proposition 2 (Nonparametric sharp bounds on the ACTE for binary outcomes). Under Assumptions 1 and 2 and when $J = 3$, the nonparametric sharp bounds on $\tau(a, a' \mid c)$ for a binary outcome can be obtained as a solution to the following linear programming problem.

$$\min_{\Phi} \quad \text{and} \quad \max_{\Phi} \quad \frac{1}{\Pr(A_i = c|D_i = 0)} \times \left\{ \sum_{a'' \in \{0,1\}} \sum_{s \in \mathcal{A}} \left(\phi_{1,0,y_{a''},s,c} - \phi_{0,1,y_{a''},s,c} \right) \right\},$$

s.t. $\phi_{y_0,y_1,y_2,s,c'} \geq 0 \; \forall \, y_0, y_1, y_2, s, c'$, $\sum_{y_0 \in \{0,1\}} \sum_{y_1 \in \{0,1\}} \sum_{y_2 \in \{0,1\}} \sum_{s \in \mathcal{A}} \sum_{c' \in \mathcal{A}} \phi_{y_0,y_1,y_2,s,c'} = 1$, $\sum_{y_0 \in \{0,1\}} \sum_{y_1 \in \{0,1\}} \sum_{y_2 \in \{0,1\}} \phi_{y_0,y_1,y_2,s,c'} \cdot \mathbf{1}\{y_{c'} = 1\} = \Pr(S_i = s, A_i = c', Y_i = 1 \mid D_i = 0) \; \forall \, s, c'$, $\sum_{y_0 \in \{0,1\}} \sum_{y_1 \in \{0,1\}} \sum_{y_2 \in \{0,1\}} \phi_{y_0,y_1,y_2,s,c'} = \Pr(S_i = s, A_i = c' \mid D_i = 0) \; \forall \, s, c'$, and $\sum_{y_0 \in \{0,1\}} \sum_{y_1 \in \{0,1\}} \sum_{y_2 \in \{0,1\}} \sum_{c' \in \mathcal{A}} \phi_{y_0,y_1,y_2,s,c'} \cdot \mathbf{1}\{y_{a''} = 1\} = \Pr(S_i = s, Y_i = 1 \mid A_i = a'', D_i = 1) \; \forall \, s, a''$, where $\Phi \equiv \left\{ \phi_{y_0,y_1,y_2,s,c} : y_0 \in \{0, 1\}, y_1 \in \{0, 1\}, y_2 \in \{0, 1\}, s \in \mathcal{A}, c \in \mathcal{A} \right\}$.

A proof is provided in Web Appendix A4. The maximization and minimization problems in Proposition 2 are standard linear programming problems which can be easily solved numerically with given data using statistical software, such as the `lpSolve` package in R. We note that, for the $\tau(a, c|c)$ case with binary outcomes, Equation (7) simplifies to

$$\begin{aligned}
&\sum_{s \in \mathcal{A}} \left(1 - \min \left\{ 1, \frac{F(0|s,a,1) - F(0|s,a,0) P(a|s,0)}{\Pr(A_i = c|S_i = s, D_i = 0)} \right\} \right) \\
&\qquad \times \Pr(S_i = s|A_i = c, D_i = 0) \\
&\qquad - \Pr(Y_i = 1|A_i = c, D_i = 0) \\
&\qquad \leq \; \tau(a, c|c) \; \leq \\
&\sum_{s \in \mathcal{A}} \left(1 - \max \left\{ 0, 1 - \frac{1 - F(0|s,a,1) + \{1 - F(0|s,a,0)\} P(a|s,0)}{\Pr(A_i = c|S_i = s, D_i = 0)} \right\} \right) \\
&\qquad \times \Pr(S_i = s|A_i = c, D_i = 0) \\
&\qquad - \Pr(Y_i = 1|A_i = c, D_i = 0),
\end{aligned}$$

which we find to numerically coincide with the linear programming bounds based on Proposition 2, as they should.

5. Sensitivity Analysis

The nonparametric bounds in Propositions 1 and 2 represent "worst-case" scenarios, in that they allow for the maximal deviation in the average potential outcomes between those subjects who merely state they would take a treatment and those who actually choose to take the treatment. In contrast, the naïve estimator given in Equation (5) relies on Assumption 3 and assumes (often demonstrably falsely) that this deviation is zero. The truth, however, lies somewhere between these two extremes.

In this section, we propose a sensitivity analysis to investigate this middle ground. Sensitivity analysis is a commonly used inferential strategy where the degree of violation of a key identification assumption is quantified via a sensitivity parameter (Rosenbaum 2002) and the consequence of this violation is then expressed and analyzed as a function of this parameter. Here, we consider a sensitivity parameter ρ_{ac} which represents the maximum absolute difference we allow to exist between the average of a potential outcome ($Y_i(a)$) among those who state a particular treatment preference ($S_i = c$) and the average of the same potential outcome among those who actually choose that treatment ($C_i = c$). Formally, ρ_{ac} is defined by the following inequality,

$$|\mathbb{E}[Y_i(a)|S_i = c] - \mathbb{E}[Y_i(a)|C_i = c]| \; \leq \; \rho_{ac},$$

which implies the following additional constraint for identification analysis under Assumptions 1 and 2,

$$\begin{aligned}&\mathbb{E}[Y_i|S_i=c,A_i=a,D_i=1]-\rho_{ac}\\&\quad\leq\ \mathbb{E}[Y_i(a)|C_i=c]\\&\quad\leq\ \mathbb{E}[Y_i|S_i=c,A_i=a,D_i=1]+\rho_{ac},\end{aligned}\tag{8}$$

for a given pair of a and $c \in \mathcal{A}$ such that $a \neq c$.

The proposed sensitivity analysis proceeds by combining the sensitivity constraint in Equation (8) with the no-assumption bounds on average choice-specific potential outcomes,

$$\begin{aligned}&\sum_{s\in\mathcal{A}}\left\{\pi^-(a|s,c)\Pr(S_i=s|A_i=c,D_i=0)\right\}\\&\quad\leq\ \mathbb{E}[Y_i(a)|C_i=c]\\&\quad\leq\ \sum_{s\in\mathcal{A}}\left\{\pi^+(a|s,c)\Pr(S_i=s|A_i=c,D_i=0)\right\},\end{aligned}$$

where $\pi^-(a|s,c)$ and $\pi^+(a|s,c)$ are the sharp bounds on $\mathbb{E}[Y_i(a) \mid S_i = s, C_i = c]$ derived in Web Appendix A3. We therefore find bounds on $\tau(a,a'|c)$ for a given pair of $(\rho_{ac}, \rho_{a'c})$ by the interval difference

$$\begin{aligned}\tau(a,a'|c)\ \in\ &\left(\left[\mathbb{E}[Y_i|S_i=c,A_i=a,D_i=1]\right.\right.\\&\left.-\rho_{ac},\ \mathbb{E}[Y_i|S_i=c,A_i=a,D_i=1]+\rho_{ac}\right]\\&\cap\left[\sum_{s\in\mathcal{A}}\left\{\pi^-(a|s,c)\Pr(S_i=s|A_i=c,D_i=0)\right\},\right.\\&\left.\left.\sum_{s\in\mathcal{A}}\left\{\pi^+(a|s,c)\Pr(S_i=s|A_i=c,D_i=0)\right\}\right]\right)\\&-\left(\left[\mathbb{E}[Y_i|S_i=c,A_i=a',D_i=1]-\rho_{a'c},\right.\right.\\&\left.\left.\mathbb{E}[Y_i|S_i=c,A_i=a',D_i=1]+\rho_{ac}\right]\right.\\&\cap\left[\sum_{s\in\mathcal{A}}\left\{\pi^-(a'|s,c)\Pr(S_i=s|A_i=c,D_i=0)\right\},\right.\\&\left.\left.\sum_{s\in\mathcal{A}}\left\{\pi^+(a'|s,c)\Pr(S_i=s|A_i=c,D_i=0)\right\}\right]\right).\end{aligned}\tag{9}$$

Again, when $a' = c$, this expression can be simplified by substituting the second interval with the observed quantity $\mathbb{E}[Y_i|A_i = c, D_i = 0]$, and the resulting bounds are now sharp for a given value of ρ_{ac}. When $a' \neq c$, we recommend setting $\rho_{ac} = \rho_{a'c} = \rho$ and conducting the analysis with respect to a single sensitivity parameter for both counterfactual outcomes for the sake of interpretability.

Before proceeding to binary outcomes, we note that a similar sensitivity approach can be applied to a simpler design with only the forced-exposure arm. However, this approach is limited in two ways. First, the sensitivity bounds will be much wider because they will consist of Equation (9) without the latter portion of each intersection. Second, because the free-choice arm is the only source of information about the feasible range of ρ_{ac}, even infinite values cannot be excluded in the simpler design when the outcome variable is unbounded. At this extreme, the sensitivity analysis becomes completely uninformative, as noted in Section 4.

For binary outcomes, the sharp sensitivity bounds can be numerically obtained for any $\tau(a,a'|c)$ and given values of ρ_{ac} and $\rho_{a'c}$ by incorporating Equation (8) into the linear programming problem in Proposition 2 as another set of linear constraints. For the special case of $J = 3$, these constraints can be written in terms of $\phi_{y_0,y_1,y_2,s,c}$ as $\sum_{y_0\in\{0,1\}}\sum_{y_1\in\{0,1\}}\sum_{y_2\in\{0,1\}}\sum_{s\in\mathcal{A}}\phi_{y_0,y_1,y_2,s,c}\mathbf{1}\{y_{a^*} = 1\} \geq \{\Pr(Y_i = 1 \mid S_i = c, A_i = a^*, D_i = 1) - \rho_{a^*c}\}\Pr(A_i = c|D_i = 0)$ and $\sum_{y_0\in\{0,1\}}\sum_{y_1\in\{0,1\}}\sum_{y_2\in\{0,1\}}\sum_{s\in\mathcal{A}}\phi_{y_0,y_1,y_2,s,c}\mathbf{1}\{y_{a^*} = 1\} \leq \{\Pr(Y_i = 1 \mid S_i = c, A_i = a^*, D_i = 1) + \rho_{a^*c}\}\Pr(A_i = c|D_i = 0)$ for given c and $a^* \in \{a, a'\}$.

6. Statistical Inference

In this section, we discuss our methods for performing statistical inference for the large-sample bounds in Sections 4 and 5. Several approaches have been proposed for inference about partially identified parameters in the literature, including the nonparametric bootstrap (e.g., Horowitz and Manski 2000). Here, we take a Bayesian approach where we obtain simulated draws from the marginal posterior distribution for the bounds on $\tau(a,a'|c)$ by Monte Carlo integration of the approximated joint posterior.

For inference about the general bounds in Proposition 1, we use the following procedure to obtain one simulated draw of the bounds, $\tau^-(a,a'|c)^*$ and $\tau^+(a,a'|c)^*$, from their joint posterior.

Algorithm 1 (Posterior simulation for the bounds in Proposition 1).

1. Draw $\boldsymbol{p} \equiv [p_s] \sim \text{Dirichlet}(\boldsymbol{n})$, where $\boldsymbol{n} \equiv [n_s] = \left[\sum_{i=1}^N \mathbf{1}\{S_i = 0\}, \ldots, \sum_{i=1}^N \mathbf{1}\{S_i = J-1\}\right]^\top$.
2. For each $s \in \mathcal{A}$:
   (a) Draw $\boldsymbol{q}_s \equiv [q_{sa}] \sim \text{Dirichlet}(\boldsymbol{n}_s^0)$, where $\boldsymbol{n}_s^0 \equiv [n_{sa}^0] = \left[\sum_{i=1}^N \mathbf{1}\{S_i = s, A_i = 0, D_i = 0\}, \ldots, \sum_{i=1}^N \mathbf{1}\{S_i = s, A_i = J-1, D_i = 0\}\right]^\top$.
   (b) For each a and $c \in \mathcal{A}$, draw a pair $[\pi^-(a|s,c), \pi^+(a|s,c)]$ from Normal$\left(\left[\begin{smallmatrix}\bar{\pi}^-\\ \bar{\pi}^+\end{smallmatrix}\right], \left[\begin{smallmatrix}V^- & C\\ C & V^+\end{smallmatrix}\right]\right)$, where expressions for $\bar{\pi}^-, \bar{\pi}^+, V^-, V^+$ and C are provided in Web Appendix A5.
3. Calculate a simulated draw of $[\tau^-(a,a'|c), \tau^+(a,a'|c)]$ as

$$\tau^-(a,a'|c)^* = \sum_{s\in\mathcal{A}}\left(\pi^-(a|s,c) - \pi^+(a'|s,c)\right)\frac{q_{sc}p_s}{\sum_{s'\in\mathcal{A}}q_{s'c}p_{s'}},$$

$$\tau^+(a,a'|c)^* = \sum_{s\in\mathcal{A}}\left(\pi^+(a|s,c) - \pi^-(a'|s,c)\right)\frac{q_{sc}p_s}{\sum_{s'\in\mathcal{A}}q_{s'c}p_{s'}}.$$

Note that for $a = c$, $\pi(a|s,c) = \pi^-(a|s,c) = \pi^+(a|s,c)$ and Step 2 will be a draw from the univariate normal distribution with mean $\bar{\pi}$ and variance V, which are also given in

Web Appendix A5. This procedure asymptotically approximates the posterior for the bounds on $\tau(a, a'|c)$ without assuming a full parametric model for the true distribution of the potential outcomes. Exact prior specifications are therefore unnecessary for these parameters; we use improper flat priors where explicit prior specifications are necessary (i.e., for $\boldsymbol{p}$ and $\boldsymbol{q}$). Details are provided in Web Appendix A5.

For the sharp bounds for binary outcomes in Proposition 2, we can apply a similar algorithm which simulates the joint posterior on the proportions of the observed strata and solves the linear programming problem using the simulated draws from the posterior. The procedure is identical to Algorithm 1 except that Steps 2 and 3 should be replaced with the following:

2. (b) For each $a \in \mathcal{A}$, draw $H_{sa} \sim \text{Beta}(\sum_{i=1}^{N} \mathbf{1}\{Y_i = 1, S_i = s, A_i = a, D_i = 1\}, \sum_{i=1}^{N} \mathbf{1}\{Y_i = 0, S_i = s, A_i = a, D_i = 1\})$ and $G_{sa} \sim \text{Beta}(\sum_{i=1}^{N} \mathbf{1}\{Y_i = 1, S_i = s, A_i = a, D_i = 0\}, \sum_{i=1}^{N} \mathbf{1}\{Y_i = 0, S_i = s, A_i = a, D_i = 0\})$.
3. Calculate a simulated draw of $[\tau^-(a, a'|c), \tau^+(a, a'|c)]$ by solving the linear programming problem in Proposition 2, with the probabilities of the observed strata in the constraints replaced with the simulated draws from their posterior ($\boldsymbol{q} \equiv [\boldsymbol{q}_s]$, $\boldsymbol{G} \equiv [G_{sa}]$, and $\boldsymbol{H} \equiv [H_{sa}]$).

Once we generate a large enough number of draws from the approximated joint posterior of $\tau^-(a, a'|c)$ and $\tau^+(a, a'|c)$, we construct $100(1-\alpha)\%$ credible intervals for the bounds. For each combination of a, a', and c, we use a numerical procedure to find the narrowest interval that entirely contains $1-\alpha$ of the draws of the bounds (i.e., $[\hat{\tau}^-(a, a'|c), \hat{\tau}^+(a, a'|c)]$ such that $\hat{\tau}^-(a, a'|c) < \tau^-(a, a'|c)^*$ and $\tau^+(a, a'|c)^* < \hat{\tau}^+(a, a'|c)$ for $100(1-\alpha)\%$ of the draws). We call these the highest posterior density (HPD) intervals for the bounds, because they contain $1-\alpha$ of the posterior probability mass defined on the bounds with the minimal width. In practice, we find that HPD intervals are virtually indistinguishable from simply taking the $\alpha/2$ and $1-\alpha/2$ quantiles of the lower and upper bounds posteriors, respectively. In a simulation study (Web Appendix A7), we find that HPD intervals generally perform better than a nonparametric bootstrap approach in finite samples in terms of frequentist coverage probability.

Finally, uncertainty estimates for the sensitivity analysis bounds in Section 5 can also be obtained following analogous procedures. For the general bounds, the key difference is that the algorithms for the sensitivity bounds are augmented to incorporate an additional set of parameters $\mathbb{E}[Y_i|S_i, A_i, D_i = 1]$ as part of the joint posterior to simulate from, which is used in the sensitivity constraint. Details are provided in Web Appendix A6.

7. Empirical Application

Now we apply the proposed methodology to the empirical example we described in Section 2.

7.1. Design and Data

In implementing the media choice experiment, we closely followed the proposed protocol as described in Section 3.1 and summarized in Figure 1. First, to measure the stated preferences over treatment options, early in the survey we asked, "If you were given the choice of the following four television programs to watch, which would you choose?" We presented each television choice (listed in Section 2) with an accompanying screenshot of the host of the show, while randomizing the order of the shows.

Subsequently, we included a "washout" period in which we asked subjects various questions not directly related to the media choice (e.g., demographics, unrelated psychological experiments). Survey researchers have long known that the order in which questions are asked can influence subsequent patterns of responses. One reason is that asking a respondent a particular question can prime certain concepts, making them salient at the time of response (Tourangeau, Rips, and Rasinski 2000). In this case, we included these filler questions to minimize the possibility that responses to the initial preference question might contaminate their subsequent media choice in the free-choice
condition.

The washout items included questions about their partisanship that we used to categorize their media preferences as pro- or counter-attitudinal. We considered Fox News "pro-attitudinal" for Republicans and "counter-attitudinal" for Democrats. MSNBC was coded in the opposite manner. After excluding subjects who were neither Democrats nor Republicans, 31% of the sample expressed a preference for pro-attitudinal media ($S_i = 0$), 12% for counter-attitudinal media ($S_i = 1$), and the remaining 56% for an entertainment show ($S_i = 2$).

Next, we randomized subjects with equal probability into the forced-exposure ($D_i = 1$) and free-choice ($D_i = 0$) conditions. We randomly assigned those in the forced-choice arm to watch pro-attitudinal media ($A_i = 0$), counter-attitudinal media ($A_i = 1$), or a randomly chosen entertainment program ($A_i = 2$), each with probability 1/3. For those in the free-choice arm, we instead asked, "Which of these programs would you like to watch now?" with the same four options presented as before. Based on their partisanship and response, we recorded the actual choice C_i as 0, 1, or 2. Here, we find that stated preferences correspond only loosely to actual choices, and that those stating a preference for entertainment were more likely to be consistent in their actual choices ($\Pr(C_i = 2|S_i = 2) = 0.91$, whereas $\Pr(C_i = 0|S_i = 0) = 0.83$ and $\Pr(C_i = 1|S_i = 1) = 0.78$ in the data). We assigned these subjects to view their choice, so that $A_i = C_i$ in the free-choice arm.

We consider two outcome variables. First, after showing the program, we asked respondents to rate the clip they watched on a number of dimensions, which we subsequently summarized into an index of sentiment toward media. The index ranged between 0 and 1 and the mean and SD were 0.61 and 0.17, respectively. Second, to gauge behavioral responses, we asked subjects how likely they would be to discuss the clip with a friend, which we recoded into a binary indicator. Overall, 62.5% of subjects were at least "somewhat likely" to discuss the viewed program (= 1) while the rest answered they were "not likely" to do so (= 0).

Table 1 summarizes the observed data from the media choice experiment. The general pattern indicates that discrepancies between stated and true preferences not only exist, but that these discrepancies are also associated with different responses to

Table 1. Summary of observed data in the media choice experiment.

Free-choice condition ($D_i = 0$)										
Stated preference (S_i)			0			1			2	
Actual choice ($C_i = A_i$)		0	1	2	0	1	2	0	1	2
Strata proportions		0.25	0.02	0.03	0.01	0.09	0.02	0.03	0.02	0.53
Outcomes (Y_i)	Sentiment toward media	0.67	0.51	0.66	0.52	0.56	0.60	0.60	0.54	0.68
	Likely to discuss	0.77	0.76	0.62	0.62	0.75	0.68	0.82	0.77	0.57
Forced-exposure condition ($D_i = 1$)										
Stated preference (S_i)			0			1			2	
Randomized treatment (A_i)		0	1	2	0	1	2	0	1	2
Strata proportions		0.10	0.11	0.11	0.04	0.05	0.05	0.20	0.18	0.16
Outcomes (Y_i)	Sentiment toward media	0.67	0.38	0.64	0.59	0.54	0.63	0.57	0.47	0.64
	Likely to discuss	0.74	0.48	0.42	0.73	0.76	0.66	0.66	0.56	0.56

NOTE: The third row in each table shows the observed proportion in each stated preference-treatment stratum. The bottom two rows in each table represent the sample averages of the two outcome variables in each stratum.

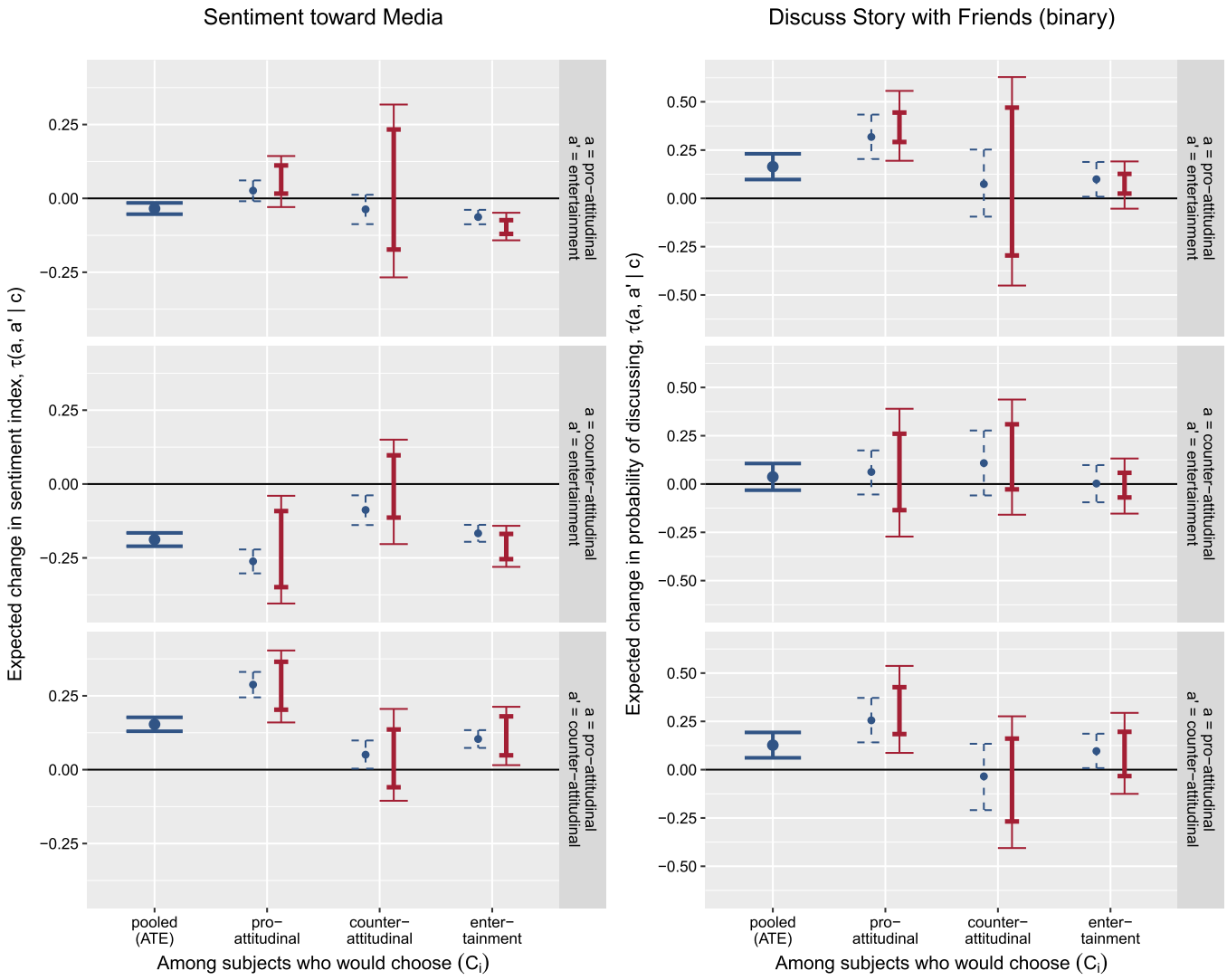

Figure 2. Estimated nonparametric bounds on the ACTE of partisan news media. Vertically stacked plots correspond to the same outcome variable. Horizontally aligned plots depict the effect of a particular change in the assigned media, that is, $\mathbb{E}[Y_i(a) - Y_i(a')|C_i = c]$. Pairs of lines correspond to the ACTE among those that would choose a given media (horizontal axis labels). Large blue points and solid thick blue error bars are pooled ATEs. Small blue points are naïve estimates, with blue dashed error bars representing 95% asymptotic confidence intervals. Solid thick red error bars are estimated bounds (posterior means) and thin error bars give 95% posterior intervals.

media. For example, among those respondents in the free-choice group who stated a preference for pro-attitudinal media and also chose a pro-attitudinal program, mean sentiment was 0.67. In contrast, responses were significantly lower (by 0.07) among free-choice subjects who stated a preference for entertainment but actually chose pro-attitudinal media.

7.2. Nonparametric Bounds

Given the evidence that stated preferences of subjects do not accurately reflect their actual choices, we now seek to bound the ACTEs using the method developed in Section 4. Figure 2 presents the resulting nonparametric bounds, along with their 95% posterior intervals obtained via the procedure proposed in Section 6. The left panel presents results for subjects' sentiment toward the media watched (Proposition 1), and the right panel presents results for whether respondents were likely to discuss the story with a friend (binary; Proposition 2). Each vertically arrayed plot depicts the effect of a particular change in the assigned media, from entertainment to pro-attitudinal (top), entertainment to counter-attitudinal (middle), and counter- to pro-attitudinal (bottom). The leftmost blue solid circle (point estimate) and error bar (95% asymptotic confidence interval) in each plot is the pooled ATE. Paired lines within each plot (thin blue and thick red) represent the estimated ACTE of that treatment among subjects that would choose pro-attitudinal media (left), counter-attitudinal media (middle), and an entertainment show (right). Small blue points are the point estimates under Assumptions 1–3, that is, the naïve estimates that assume the ignorability of the discrepancy between stated preferences and actual choices. Blue dashed error bars are 95% asymptotic confidence intervals. Solid red error bars are nonparametric bounds on ACTEs under Assumptions 1 and 2 alone, with thick lines representing estimated bounds (posterior means) and thin lines representing posterior intervals.

For example, consider the middle bars in the center left plot. Here, blue dashed estimates show that, even among subjects that state a preference for counter-attitudinal media, this media results in more negative sentiment than entertainment—while small, the naïve estimate is negative and statistically significant at the 95% confidence level. In contrast, the sharp bounds, centered directly on zero, show that this result may be misleading for the group that would actually choose counter-attitudinal media, because inconsistency in stated and true preferences may be systematically correlated with responses. Indeed, in Section 7.3, we will show that it is highly sensitive to assumptions about the informativeness of the stated preference. The greatest source of this discrepancy is that for counter-attitudinal media, stated preferences are particularly inconsistent with actual choices. In the free-choice condition, over 20% of subjects stating this preference went on to choose other media.

We now briefly discuss the remaining estimates in the left panel of Figure 2, starting with the top left and proceeding clockwise. In the top plot, all bounds agree with naïve estimates: differences in sentiment toward pro-attitudinal media and entertainment are indistinguishable, except for a small adverse reaction among those with a true preference for entertainment (top right). These same subjects have a significant and seemingly larger adverse reaction to counter-attitudinal media (center right), and the difference between pro- and counter-attitudinal media among this group is statistically significant (lower right). Among units that would choose counter-attitudinal media, naïve estimates suggest a significantly more positive reaction to pro- versus counter-attitudinal media (lower middle), but these results again implicitly rest on strong assumptions about the informativeness of stated preferences. Not surprisingly, those who would choose pro-attitudinal media react more positively toward it than toward counter-attitudinal media (lower left). Finally, estimated bounds appear to support the naïve estimate that those who would choose pro-attitudinal media have a negative response to counter-attitudinal media (vs. entertainment, center left) and these bounds are statistically distinct from zero.

Finally, we present nonparametric sharp bounds for the binary outcome of whether subjects are likely to discuss the story with a friend. As explained in Section 4.2, these are the narrowest possible bounds that can be found with the available information. We discuss statistically significant results only. Among units that would choose pro-attitudinal media, bounds validate the naïve estimate that this media has a large effect on the dissemination of information, both relative to entertainment (top left) and relative to counter-attitudinal media (bottom left). Naïve estimates suggest a similar but smaller pattern of effects for those who would choose entertainment. However, the estimated bounds are, respectively, consistent with the naïve estimate in sign but statistically inconclusive (vs. entertainment, top right) and entirely inconclusive (vs. counter-attitudinal media, bottom right).

7.3. Sensitivity Analysis

Next, we apply the sensitivity analysis developed in Section 5 and show how the bounds become tighter as we allow less difference between the average potential outcomes conditional on a stated preference versus actual choice (ρ). For illustration, we focus on the sentiment index. The results are presented in Figure 3. The results for the binary discussion indicator are in Figure A.1 in the Web Appendix.

Using bounds on mean choice-specific potential outcomes (not presented), we find that the estimated maximal difference for any strata is 0.18, approximately one SD in the outcome variable. Thus, in Figure 3, estimated sensitivity results have converged to the estimated bounds at or below this level of ρ. For most strata, differences above 0.1 can be ruled out. For some ACTEs, sensitivity results are not shown for low values of ρ because in this region, it becomes impossible to simultaneously satisfy the sensitivity constraints implied by ρ and the naïve results, on the one hand, and the bounding constraints, on the other.

For illustration we focus on the middle plot in the center row of Figure 3. Neglecting sampling error, the true value of ρ here should lie somewhere in $[0.02, 0.18]$. The naïve estimates suggest that counter-attitudinal media negatively affects media sentiment (relative to entertainment) even among those who would choose counter-attitudinal media (middle plot), somewhat surprisingly. However, the upper bound is statistically indistinguishable from zero when Assumption 3 is even slightly relaxed by $\rho = 0.03$. Estimated bounds include zero for values of $\rho > 0.07$, less than half of a SD in the observed outcome variable.

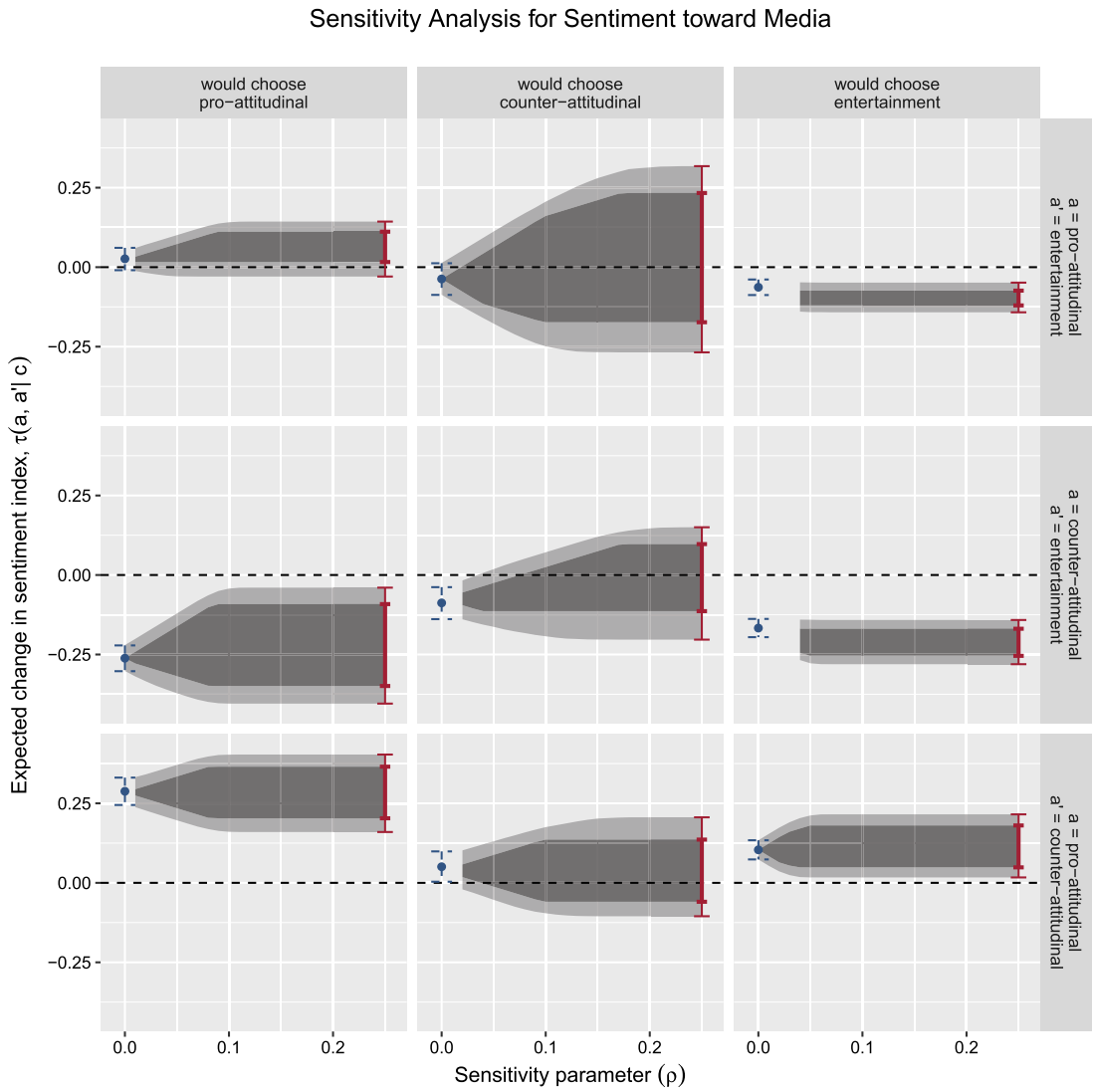

Figure 3. Sensitivity analysis for the ACTE of partisan news media. The plots correspond to the left panel of Figure 2. On the left side of each plot, a blue point and error bars represent the naïve estimate and 95% asymptotic confidence intervals, respectively. On the right side, thick red error bars represent no-assumption bounds and thin red error bars represent 95% posterior intervals. The dark shaded region between these depicts how bounds grow narrower as additional information from the naïve estimates are incorporated ($\rho_{ac} = \rho_{a'c} = \rho$ grows small). Lightly shaded regions are 95% posterior regions for sensitivity results.

8. Simulations

To evaluate the finite-sample performance of our procedure in a naturalistic setting, we conduct Monte Carlo simulations that are based closely on the observed data from our empirical application. We design our simulation study to consider two challenges that regularly arise in PPTs: the discrepancy between stated preferences and actual choices, and the possibility that choices are related to potential outcomes in an unobservable way. We quantify these complications via a "choice divergence" (CD) parameter and an "outcome divergence" (OD) parameter, respectively, and examine how bias and coverage rates are affected as they become more severe.

8.1. Divergence of Actual Choices From Stated Preferences

In our first set of simulations, we begin by generating a hypothetical population for which the observable margins are identical to those of our empirical sample. For the unobserved variables, such as $Y_i(a)$ for $C_i \neq a$, we make the most generous assumption possible—that potential outcomes among the unobserved choice-based subgroups are identical in distribution—and generate the values accordingly. This forms our baseline simulation data, in which Assumption 3 is satisfied and therefore the naïve estimator for the ACTE should perform well.

We then introduce the CD parameter, which takes on values in $[0, 1]$ and captures the informativeness of the stated preferences. That is, we regenerate the actual choice C_i^* such that $\Pr(C_i^* = c|S_i = s) = (1 - \text{CD})\Pr(C_i = c|S_i = s) + \text{CD}/J$, where $C_i \in \{1, \ldots, J\}$ for each c and s. When CD = 0, the joint distribution of C_i^* and S_i is identical to that of the experimental sample (i.e., with 87% overall agreement between stated preferences and actual choices). When CD = 1, stated preferences are entirely uninformative, and the choice probabilities are 1/3 for every treatment.

For each value at which the CD parameter is set (0, 1/3, 2/3, and 1), we draw 500 sample datasets of size 3000. We focus on $\tau(0, 2|0)$ for illustration. Naïve ACTE estimates and bounds estimates are computed in each sample dataset, averaged, and compared to the population ACTE and bounds to assess bias. The results of these simulations are collected in Table 2 (left panel) and show that the bias of the naïve estimator increases as stated preferences diverge more from actual choices. At the extreme, the bias of the naïve estimate is well over double the SE of the naïve estimate in a particular sample dataset (roughly 0.016 at this sample size). In contrast, ACTE bounds estimates are centered almost exactly on their population analogue, and bias is unaffected by the CD parameter.

For each sample dataset, we also implement the procedure described in Section 6 to compute 95% posterior intervals on the bounds. The proportion of intervals that fully contain the population bounds is reported in Table 2 (right panel) for sample sizes of ranging from 500 to 50,000 units. We find that when stated preferences are highly divergent from actual choices (CD is high), coverage rates are less than nominal with small sample sizes, but they become indistinguishable from 95% as the number of observations increase. We note that when the bounds are wide relative to their credible interval, coverage of the population bounds at level $1-\alpha$ corresponds to a worst-case coverage of the population ACTE at level $1-\alpha/2$. In Web Appendix A7, the proposed method is shown to outperform both alternative bootstrap-based approach and an extension to the parametric model of Long, Little, and Lin (2008).

8.2. Divergence of Outcomes Between Choice Groups

Next, we define the OD parameter, also in [0, 1], which controls the distributions of the unobservable potential outcomes, $\Pr(Y_i(a) \leq y|S_i = s, C_i = c, D_i = 0)$ when $c \neq a$. Specifically, we generate missing potential outcomes for the free-choice group such that $\Pr(Y_i(a) \leq y|S_i = s, C_i = c, D_i = 0) = (1-\text{OD})\{F(y|s,c,1) - F(y|s,c,0)P(a|s,0)\}/\{1-P(a|s,0)\} + \text{OD} \cdot \mathbf{1}\{F(y|s,a,1) - F(y|s,a,0)P(a|s,0) < P(c|s,0)\}\{F(y|s,a,1) - F(y|s,a,0)P(a|s,0)\}/P(c|s,0)$ for a particular $c \neq a$, which then fixes the distribution of potential outcomes in the remaining choice-based subgroup. All values of the OD parameter produce observationally equivalent simulation populations, but the population ACTE differs depending on its setting. When OD = 1, the population ACTE is the most extreme effect consistent with the observable margins, lying on the endpoint of the population bounds. In contrast, OD = 0 produces the least extreme population satisfying the same observed margins, with $\Pr(Y_i(a) \leq y|S_i = s, C_i = c)$ identical to $\Pr(Y_i(a) \leq y|S_i = s, C_i = c')$. In these simulations, CD is set to zero.

Table 3 (left panel) presents the bias of naïve and bounds estimates at varying levels of the OD parameter, with sample sizes of 3000. Once again, results demonstrate that while naïve estimates are accurate in a best-case simulation, bias is on the order of one SE of any particular naïve estimate when assumptions are violated even moderately (OD between 0.33 and 0.67). Bias of the proposed bounds procedure remains negligible for all OD values. Table 3 (right panel) shows that coverage levels are as expected over a wide range for OD and N, or perhaps slightly conservative at smaller sample sizes.

Table 2. Estimated bias and coverage rates for various CD values.

	CD = 0.00	CD = 0.33	CD = 0.67	CD = 1.00
Naïve	0.002	0.011	0.023	0.038
Min	−0.001	−0.001	−0.001	0.000
Max	−0.001	−0.001	−0.001	−0.001

N	CD = 0.00	CD = 0.33	CD = 0.67	CD = 1.00
500	0.959	0.944	0.914	0.926
1000	0.956	0.926	0.930	0.924
3000	0.946	0.936	0.916	0.946
10,000	0.938	0.938	0.934	0.930
50,000	0.946	0.944	0.940	0.948

NOTE: Results for each CD value are based on 500 sample datasets. The bias results for the naïve and the bounds lower/upper endpoint estimators (left) are based on datasets of size $N = 3000$. The bounds posterior interval coverage rates (right) have Monte Carlo SEs of approximately 0.01.

Table 3. Estimated bias and coverage rates for various OD values, holding CD at 0.

	OD = 0.00	OD = 0.33	OD = 0.67	OD = 1.00
Naïve	0.002	0.011	0.020	0.030
Min	−0.001	0.001	0.001	0.001
Max	−0.001	−0.002	−0.002	−0.001

N	OD = 0.00	OD = 0.33	OD = 0.67	OD = 1.00
500	0.959	0.967	0.970	0.962
1000	0.956	0.966	0.971	0.968
3000	0.946	0.954	0.950	0.968
10,000	0.938	0.954	0.958	0.966
50,000	0.946	0.954	0.954	0.946

NOTE: The left table presents bias results for the naïve and the bounds lower/upper endpoint estimators, and the right table presents bounds posterior interval coverage rates. These results are based on the procedure of Table 2, which presents the same quantities for varying CD values.

9. Concluding Remarks

Scholars of social and medical sciences have long sought to enhance the external validity of randomized experiments through various means. Medical researchers have often adopted PPTs to incorporate the preferences of experimental subjects over treatment options into their study designs, thereby tackling the question of what impact treatments have on the kinds of people who would actually take them if they were allowed to choose. However, systematic analysis of causal and statistical properties of PPTs has only just begun. In particular, researchers have largely neglected the potential discrepancy between subjects' stated and revealed preferences in the existing literature.

In this article, we seek to address the challenge of improving external validity via a new experimental design for PPTs. The proposed design involves measurement of both stated preferences and actual choices as well as randomization into the standard RCT or a free-choice condition. The methodology we develop systematically addresses the potential inferential threat caused by nonignorable differences between stated and revealed preferences, using both nonparametric identification analysis and sensitivity analysis. As we illustrate in an original empirical example where we use the proposed framework, our method enables inference on a causal quantity of interest that captures the heterogeneity in treatment effects across revealed preferences without relying on the assumption of ignorable measurement error. We provide open-source software, `ppt`, which implements the proposed methodology.

Future statistical work on PPTs should investigate the consequence of noncompliance and differential attrition on the estimation of ACTEs, among other inferential challenges left unaddressed by the current article. An important motivation for PPTs in medical research is the concern that a patient who strongly prefers one treatment option may not follow experimental protocols and cross over to another treatment arm or out of the study, damaging the internal validity of the experiment. Forcing patients into treatment options against their preferences may also be considered unethical in some applications (Lambert and Wood 2000). One natural direction for future research is, therefore, to incorporate such complications under the current framework.

Supplementary Materials

The online supplementary materials contain the Web Appendices referred to in the main text. They consist of the following sections: (1) Observable Implications of Assumption 3; (2) Derivation of Equation 4; (3) Proof of Proposition 1; (4) Proof of Proposition 2; (5) Statistical Inference for the Bounds; (6) Statistical Inference for the Sensitivity Analysis; and (7) Additional Simulation Results. (UASA_A_1585248_SM6304.zip)

Acknowledgments

We are grateful to Donald Green and David Nickerson for their helpful comments and suggestions. We also thank participants at the 2014 Society for Political Methodology Summer Meeting, MacMillan-CSAP Workshop at Yale University, and the 3rd ISAT/TSE Conference in Political Economy/Political Economy in Toulouse for their feedback.

Funding

We acknowledge financial support from the National Science Foundation (SES-1528487 and Graduate Research Fellowship under grant no. 1122374).

References

Arceneaux, K., Johnson, M., and Murphy, C. (2012), "Polarized Political Communication, Oppositional Media Hostility, and Selective Exposure," *Journal of Politics*, 74, 174–186. [1,2,4,5,6]

Balke, A. (1995), "Probabilistic Counterfactuals: Semantics, Computation, and Applications," Ph.D. thesis, Computer Science Department, University of California, Los Angeles. [7]

Balke, A., and Pearl, J. (1997), "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176. [7]

Brewin, C. R., and Bradley, C. (1989), "Patient Preferences and Randomised Clinical Trials," *BMJ: British Medical Journal*, 299, 313. [3]

Brown, N. R., and Sinclair, R. C. (1999), "Estimating Number of Lifetime Sexual Partners: Men and Women Do It Differently," *Journal of Sex Research*, 36, 292–297. [2]

Campbell, A., Converse, P. E., Miller, W., and Stokes, D. (1960), *The American Voter*, Chicago: University of Chicago Press. [2]

Clausen, A. R. (1968), "Response Validity: Vote Report," *Public Opinion Quarterly*, 32, 588–606. [2]

Frangakis, C. E., and Rubin, D. B. (2002), "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29. [7]

Gaines, B. J., and Kuklinski, J. H. (2011), "Experimental Estimation of Heterogeneous Treatment Effects Related to Self-Selection," *American Journal of Political Science*, 55, 724–736. [1,3,5]

Gallup (2014), "Media Use and Evaluation," Technical Report, Gallup Historical Trends, available at *http://www.gallup.com/poll/1663/media-use-evaluation.aspx*. [2]

Hamilton, J. T. (2005), "The Market and the Media," in *The Press*, eds. G. Overholser and K. H. Jamieson, Oxford: Oxford University Press. [2]

Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, X.-H. (2000), "Assessing the Effect of an Influenza Vaccine in an Encouragement Design," *Biostatistics*, 1, 69–88. [4]

Horowitz, J. L., and Manski, C. F. (2000), "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data," *Journal of the America Statistical Association*, 95, 77–84. [8]

Howard, L., and Thornicroft, G. (2006), "Patient Preference Randomised Controlled Trials in Mental Health Research," *The British Journal of Psychiatry*, 188, 303–304. [1]

Hser, Y.-i., Maglione, M., and Boyle, K. (1999), "Validity of Self-Report of Drug Use Among STD Patients, ER Patients, and Arrestees," *American Journal of Drug and Alcohol Abuse*, 25, 81–91. [2]

Imai, K., Tingley, D., and Yamamoto, T. (2013), "Experimental Designs for Identifying Causal Mechanisms" (with discussions), *Journal of the Royal Statistical Society*, Series A, 176, 5–51. [7]

Iyengar, S., and Hahn, K. S. (2009), "Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use," *Journal of Communication*, 59, 19–39. [2]

Kim, Y. M. (2009), "Issue Publics in the New Information Environment: Selectivity, Domain Specificity, and Extremity," *Communication Research*, 36, 254–284. [2]

King, M., Nazareth, I., Lampe, F., Bower, P., Chandler, M., Morou, M., Sibbald, B., and Lai, R. (2005), "Impact of Participant and Physician Intervention Preferences on Randomized Trials: A Systematic Review," *Journal of the American Medical Association*, 293, 1089–1099. [1,3]

Kowalski, C. J., and Mrdjenovich, A. J. (2013), "Patient Preference Clinical Trials: Why and When They Will Sometimes Be Preferred," *Perspectives in Biology and Medicine*, 56, 18–35. [3]

Ladd, J. M. (2012), *Why Americans Hate the Media and How It Matters*, Princeton: Princeton University Press. [2]

Lambert, M. F., and Wood, J. (2000), "Incorporating Patient Preferences Into Randomized Trials," *Journal of Clinical Epidemiology*, 53, 163–166. [14]

Levendusky, M. S. (2013), "Why Do Partisan Media Polarize Viewers?," *American Journal of Political Science*, 57, 611–623. [2,3]

Long, Q., Little, R. J., and Lin, X. (2008), "Causal Inference in Hybrid Intervention Trials Involving Treatment Choice," *Journal of the American Statistical Association*, 103, 474–484. [2,3,5,13]

Manski, C. F. (1995), *Identification Problems in the Social Sciences*, Cambridge, MA: Harvard University Press. [6]

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9 (translated in 1990)," *Statistical Science*, 5, 465–480. [1]

Payne, G. J. (2010), "The Bradley Effect: Mediated Reality of Race and Politics in the 2008 U.S. Presidential Election," *American Behavior Scientist*, 54, 417–435. [2]

Prior, M. (2007), *Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections*, Cambridge: Cambridge University Press. [2]

——— (2009), "The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure," *Public Opinion Quarterly*, 73, 1–14. [2]

Rogers, T., and Aida, M. (2013), "Vote Self-Prediction Hardly Predicts Who Will Vote, and Is (Misleadingly) Unbiased," *American Politics Research*, 42, 503–528. [2]

Rosenbaum, P. R. (2002), *Observational Studies* (2nd ed.), New York: Springer-Verlag. [7]

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701. [1]

——— (1990), Comments on "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9" by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed, *Statistical Science*, 5, 472–480. [3]

Schmoor, C., Olschewski, M., and Schumacher, M. (1996), "Randomized and Non-randomized Patients in Clinical Trials: Experiences With Comprehensive Cohort Studies," *Statistics in Medicine*, 15, 263–271. [3]

Stroud, N. J. (2011), *The Politics of News Choice*, Oxford: Oxford University Press. [2]

Torgerson, D. J., Klaber-Moffett, J., and Russell, I. T. (1996), "Patient Preferences in Randomised Trials: Threat or Opportunity?," *Journal of Health Services Research & Policy*, 1, 194–197. [3]

Tourangeau, R. (1999), "Remember What Happened: Memory Errors and Survey Reports," in *Memory: The Science of Self Report: Implications for Research and Practice*, eds. A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, and V. S. Cain, Hove: Psychology Press. [3]

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000), *The Psychology of Survey Response*, Cambridge, UK: Cambridge University Press. [9]

Yamamoto, T. (2012), "Understanding the Past: Statistical Analysis of Causal Attribution," *American Journal of Political Science*, 56, 237–256. [7]

Supplementary Materials

A.1 Observable Implications of Assumption 3

In this section, we derive the two observable implications of Assumption 3 described in Section 3.3. First, Assumption 3 implies,

$$\mathbb{E}[Y_i(a)|C_i = a] \;=\; \mathbb{E}[Y_i(a)|S_i = a], \tag{10}$$

for all $a \in \mathcal{A}$. This relationship directly implies equation (2) under Assumptions 1 and 2. Second, note that equation (10) also implies,

$$\begin{aligned}
\mathbb{E}[Y_i(a)|C_i = a] \;=\; & \mathbb{E}[Y_i(a)|C_i = a, S_i = a]\Pr(C_i = a|S_i = a) \\
& + \mathbb{E}[Y_i(a)|C_i \neq a, S_i = a]\Pr(C_i \neq a|S_i = a) \\
\Leftrightarrow \quad \mathbb{E}[Y_i(a)|C_i \neq a, S_i = a] \;=\; & \frac{\mathbb{E}[Y_i|C_i = a, D_i = 0] - \mathbb{E}[Y_i|C_i = S_i = a, D_i = 0]\Pr(C_i = a|S_i = a, D_i = 0)}{1 - \Pr(C_i = a|S_i = a, D_i = 0)}
\end{aligned}$$

for all $a \in \mathcal{A}$. Setting the unobserved term in the left-hand side to its theoretical maximum and minimum yields equation (3).

A.2 Derivation of Equation (4)

First, consider $\mathbb{E}[Y_i(a)|C_i = c]$. Assumptions 1 and 2 imply $\Pr(C_i = c, S_i = s) = \Pr(C_i = c, S_i = s|D_i = 0)$, $\mathbb{E}[Y_i(c)|C_i = c, S_i = s] \;=\; \mathbb{E}[Y_i|C_i = c, S_i = s, D_i = 0]$, $\mathbb{E}[Y_i(a)] = \mathbb{E}[Y_i|A_i = a, D_i = 1]$, and $\mathbb{E}[Y_i(a)|S_i = s] = \mathbb{E}[Y_i|S_i = s, A_i = a, D_i = 1]$. Now, note that

$$\mathbb{E}[Y_i|A_i = a, D_i = 1] \;=\; \mathbb{E}[Y_i(a)] \;=\; \sum_{c'=0}^{J-1} \mathbb{E}[Y_i(a)|C_i = c']\Pr(C_i = c'),$$

by Assumptions 1, 2 and the law of total expectation. Substituting observed outcomes from the free-choice group and rearranging terms, we have

$$\mathbb{E}[Y_i(a)|C_i = c] \;=\; \frac{1}{\Pr(C_i = c|D_i = 0)} \left\{ \begin{array}{l} \mathbb{E}[Y_i|A_i = a, D_i = 1] \\ -\mathbb{E}[Y_i|C_i = a, D_i = 0]\Pr(C_i = a|D_i = 0) \\ -\sum_{c' \notin \{a,c\}} \mathbb{E}[Y_i(a)|C_i = c']\Pr(C_i = c'|D_i = 0) \end{array} \right\}$$

because of Assumptions 1 and 2. By the same token,

$$\mathbb{E}[Y_i(a')|C_i=c] = \frac{1}{\Pr(C_i=c|D_i=0)}\left\{\begin{array}{l}\mathbb{E}[Y_i|A_i=a',D_i=1] \\ -\mathbb{E}[Y_i|C_i=a',D_i=0]\Pr(C_i=a|D_i=0) \\ -\sum_{c'\notin\{a',c\}}\mathbb{E}[Y_i(a')|C_i=c']\Pr(C_i=c'|D_i=0)\end{array}\right\}$$

The quantity of interest is therefore

$$\begin{aligned}\tau(a,a'|c) = & \frac{1}{\Pr(C_i=c|D_i=0)}\left\{\begin{array}{l}\mathbb{E}[Y_i|A_i=a,D_i=1] \\ -\mathbb{E}[Y_i|C_i=a,D_i=0]\Pr(C_i=a|D_i=0) \\ -\sum_{c'\notin\{a,c\}}\mathbb{E}[Y_i(a)|C_i=c']\Pr(C_i=c'|D_i=0)\end{array}\right\} \\ & - \frac{1}{\Pr(C_i=c|D_i=0)}\left\{\begin{array}{l}\mathbb{E}[Y_i|A_i=a',D_i=1] \\ -\mathbb{E}[Y_i|C_i=a',D_i=0]\Pr(C_i=a'|D_i=0) \\ -\sum_{c'\notin\{a',c\}}\mathbb{E}[Y_i(a')|C_i=c']\Pr(C_i=c'|D_i=0)\end{array}\right\}\end{aligned}$$

for any a, a' and c. Thus, under Assumptions 1 and 2, we have $2(J-2)$ terms that remain unidentified when $a \neq a' \neq c$. When $a'=c$, the above simplifies to

$$\begin{aligned}\tau(a,c|c) &= \mathbb{E}[Y_i(a)|C_i=c] - \mathbb{E}[Y_i|C_i=c,D_i=0] \\ &= \frac{1}{\Pr(C_i=c|D_i=0)}\left\{\begin{array}{l}\mathbb{E}[Y_i|A_i=a,D_i=1] \\ -\mathbb{E}[Y_i|C_i=a,D_i=0]\Pr(C_i=a|D_i=0) \\ -\sum_{c'\notin\{a,c\}}\mathbb{E}[Y_i(a)|C_i=c']\Pr(C_i=c'|D_i=0)\end{array}\right\} \\ &\quad - \mathbb{E}[Y_i|C_i=c,D_i=0]\end{aligned}$$

and $J-2$ terms remain unidentified.

A.3 Proof of Proposition 1

We begin by establishing several lemmas.

Lemma .1 *Let $\Gamma_a(y,c|s,a) = \Pr(Y_i(a) \leq y, C_i \leq c | S_i = s, C_i \neq a)$. Under Assumptions 1 and 2, the sharp upper and lower bounds on $\Gamma_a(y,c|s,a)$, denoted by $\Gamma_a^+(y,c|s,a)$ and $\Gamma_a^-(y,c|s,a)$ respectively, are identified as follows.*

$$\begin{aligned}\Gamma_a^+(y,c|s,a) &= \min\left\{H(c|s,a,0),\ \frac{F(y|s,a,1)-F(y|s,a,0)P(a|s,0)}{1-P(a|s,0)}\right\}, \\ \Gamma_a^-(y,c|s,a) &= \max\left\{0,\ H(c|s,a,0)+\frac{F(y|s,a,1)-F(y|s,a,0)P(a|s,0)}{1-P(a|s,0)}-1\right\},\end{aligned}$$

for $y \in \mathcal{Y}$, $a, c, s \in \mathcal{A}$ and $d \in \{0, 1\}$, where $H(c|s, a, 0) = \Pr(A_i \leq c|S_i = s, A_i \neq a, D_i = 0)$ and $F(y|s, a, d)$ and $P(a|s, 0)$ are as defined in Proposition 1.

Proof. By the Fréchet-Hoeffding theorem, the sharp upper and lower bounds of the bivariate joint distribution function $\Gamma_a(y, c|s, a)$ are given by,

$$\Gamma_a^+(y, c|s, a) = \min\left\{\Gamma_a(\infty, c|s, a), \Gamma_a(y, \infty|s, a)\right\}, \tag{11}$$

$$\Gamma_a^-(y, c|s, a) = \max\left\{0, \Gamma_a(\infty, c|s, a) + \Gamma_a(y, \infty|s, a) - 1\right\}. \tag{12}$$

Under Assumption 1, $\Gamma_a(\infty, c|s, a) = \Pr(C_i \leq c|S_i = s, C_i \neq a) = \Pr(A_i \leq c|S_i = s, A_i \neq a, D_i = 0) = H(c|s, a, 0)$ for any c, s and $a \in \mathcal{A}$. Under Assumptions 1 and 2, we have

$$\begin{aligned}
\Gamma_a(y, \infty|s, a) &= \Pr(Y_i(a) \leq y|S_i = s, C_i \neq a) \\
&= \frac{\Pr(Y_i(a) \leq y|S_i = s) - \Pr(Y_i(a) \leq y, C_i = a|S_i = s)}{\Pr(C_i \neq a|S_i = s)} \\
&= \frac{\Pr(Y_i(a) \leq y|S_i = s) - \Pr(Y_i(a) \leq y|C_i = a, S_i = s)\Pr(C_i = a|S_i = s)}{1 - \Pr(C_i = a|S_i = s)} \\
&= \frac{F(y|s, a, 1) - F(y|s, a, 0)P(a|s, 0)}{1 - P(a|s, 0)},
\end{aligned}$$

for any a and $s \in \mathcal{A}$. Substituting these to equations (11) and (12) yields the results in Lemma .1. ∎

Lemma .2 *Let A_i^*, C_i^* and S_i^* be reordered versions of A_i, C_i and S_i, respectively, such that $C_i^* = 0$ iff $C_i = c$ (and likewise for A_i^* and S_i^*). Then, the resulting sharp bounds on $\Gamma_a(y, c \mid s, a) - \Gamma_a(y, c-1 \mid s, a)$ are also the sharp bounds on $\Gamma_a^*(y, 0 \mid s, a) - \Gamma_a^*(y, -1 \mid s, a)$ for any y and $c \in \mathcal{A}$, where $\Gamma_a^*(y, c \mid s, a) = \Pr(Y_i(a) \leq y, C_i^* \leq c|S_i = s, C_i \neq a)$.*

Proof. First, consider the sharp bounds on $\Gamma_a(y, c) - \Gamma_a(y, c-1)$. In addition to the Fréchet-Hoeffding constraints on its constituent parts,

$$\begin{aligned}
\Gamma_a(y, c|s, a) &\in \left[\Gamma_a^-(y, c|s, a), \Gamma_a^+(y, c|s, a)\right] \\
\Gamma_a(y, c-1|s, a) &\in \left[\Gamma_a^-(y, c-1|s, a), \Gamma_a^+(y, c-1|s, a)\right],
\end{aligned}$$

the increase in cumulative probability from $c-1$ to c is also subject to

$$\Gamma_a(y,c) - \Gamma_a(y,c-1) \quad \in \quad [0, \Gamma_a(\infty,c) - \Gamma_a(\infty,c-1)]\,.$$

The combination of these constraints yields

$$\begin{aligned}
&\Gamma_a(y,c|s,a)\Gamma_a(y,c-1|s,a) \\
&\quad \in [0, \Gamma_a(\infty,c) - \Gamma_a(\infty,c-1)] \bigcup \\
&\qquad \left(\left[\Gamma_a^{*-}(y,c|s,a), \Gamma_a^{*+}(y,c|s,a) \right] - \left[\Gamma_a^{*-}(y,c-1|s,a), \Gamma_a^{*+}(y,c-1|s,a) \right] \right) \\
&\quad \in \left[\max \left\{ \begin{array}{l} 0, \\ \max \left\{ \begin{array}{l} 0, \\ \Gamma_a(\infty,c|s,a) + \Gamma_a(y,\infty|s,a) - 1 \end{array} \right\} - \min \left\{ \begin{array}{l} \Gamma_a(\infty,c-1|s,a), \\ \Gamma_a(y,\infty|s,a) \end{array} \right\} \end{array} \right\}, \right. \\
&\qquad \left. \min \left\{ \begin{array}{l} \Gamma_a(\infty,c) - \Gamma_a(\infty,c-1), \\ \min \left\{ \begin{array}{l} \Gamma_a(\infty,c|s,a), \\ \Gamma_a(y,\infty|s,a) \end{array} \right\} - \max \left\{ \begin{array}{l} 0, \\ \Gamma_a(\infty,c-1|s,a) + \Gamma_a(y,\infty|s,a) - 1 \end{array} \right\} \end{array} \right\} \right]
\end{aligned}$$

Next, consider the sharp bounds on $\Gamma_a^*(y,0 \mid s,a) - \Gamma_a^*(y,-1 \mid s,a)$. Because -1 lies below the lowest possible value of C_i^*, $\Gamma_a^*(y,-1 \mid s,a)$ is necessarily zero, and bounds on the difference reduce to bounds on $\Gamma_a^*(y,0 \mid s,a)$,

$$\begin{aligned}
\Gamma_a^*(y,0|s,a) \quad &\in \quad \left[\Gamma_a^{*-}(y,0|s,a), \Gamma_a^{*+}(y,0|s,a) \right] \\
&\in \quad \left[\max\left\{0, \Gamma_a^*(\infty,0|s,a) + \Gamma_a^*(y,\infty|s,a) - 1\right\}, \min\left\{\Gamma_a^*(\infty,0|s,a), \Gamma_a^*(y,\infty|s,a)\right\} \right] \\
&\in \quad \left[\max\left\{0, \Pr(A_i = c \mid S_i = s, A_i \neq a, D_i = 0) + \Gamma_a(y,\infty|s,a) - 1\right\}, \right. \\
&\qquad\qquad \left. \min\left\{\Pr(A_i = c \mid S_i = s, A_i \neq a, D_i = 0), \Gamma_a(y,\infty|s,a)\right\} \right] \\
&\in \quad \left[\max\left\{0, \Gamma_a(\infty,c|s,a) - \Gamma_a(\infty,c-1|s,a) + \Gamma_a(y,\infty|s,a) - 1\right\}, \right. \\
&\qquad\qquad \left. \min\left\{\Gamma_a(\infty,c|s,a) - \Gamma_a(\infty,c-1|s,a), \Gamma_a(y,\infty|s,a)\right\} \right].
\end{aligned}$$

We now show that the upper bound on $\Gamma_a(y,c) - \Gamma_a(y,c-1)$ is identical to the upper bound on

$\Gamma_a^*(y, 0 \mid s, a) - \Gamma_a^*(y, -1 \mid s, a)$ in each of the following four possible cases. (1) $\Gamma_a(\infty, c|s, a) \leq \Gamma_a(y, \infty|s, a)$ and $0 \geq \Gamma_a(\infty, c-1|s, a)+\Gamma_a(y, \infty|s, a)-1$. The upper bound on $\Gamma_a(y, c)-\Gamma_a(y, c-1)$ reduces to $\min\{\Gamma_a(\infty, c) - \Gamma_a(\infty, c-1), \Gamma_a(\infty, c|s, a) - \max\{0, \Gamma_a(\infty, c-1|s, a) + \Gamma_a(y, \infty|s, a) - 1\}\}$. This implies $\Gamma_a(\infty, c|s, a)-\Gamma_a(\infty, c-1|s, a) \leq \Gamma_a(y, \infty|s, a)$, and so the upper bound on $\Gamma_a^*(y, 0 \mid s, a)$ becomes $\Gamma_a(\infty, c|s, a) - \Gamma_a(\infty, c-1|s, a)$. Since $0 \geq \Gamma_a(\infty, c-1|s, a) + \Gamma_a(y, \infty|s, a) - 1$, the upper bound on $\Gamma_a(y, c) - \Gamma_a(y, c-1)$ further reduces to $\min\{\Gamma_a(\infty, c) - \Gamma_a(\infty, c-1), \Gamma_a(\infty, c|s, a)\} = \Gamma_a(\infty, c) - \Gamma_a(\infty, c-1)$, which is identical to the upper bound on $\Gamma_a^*(y, 0 \mid s, a)$. (2) $\Gamma_a(\infty, c|s, a) \leq \Gamma_a(y, \infty|s, a)$ and $0 < \Gamma_a(\infty, c-1|s, a)+\Gamma_a(y, \infty|s, a)-1$. The upper bound on $\Gamma_a(y, c)-\Gamma_a(y, c-1)$ becomes $\min\{\Gamma_a(\infty, c) - \Gamma_a(\infty, c-1), \Gamma_a(\infty, c|s, a) - (\Gamma_a(\infty, c-1|s, a) + \Gamma_a(y, \infty|s, a) - 1)\} = \Gamma_a(\infty, c) - \Gamma_a(\infty, c-1)$, since $1 - \Gamma_a(y, \infty|s, a) > 0$. This is again identical to the upper bound on $\Gamma_a^*(y, 0 \mid s, a)$. (3) $\Gamma_a(\infty, c|s, a) > \Gamma_a(y, \infty|s, a)$ and $0 \geq \Gamma_a(\infty, c-1|s, a) + \Gamma_a(y, \infty|s, a) - 1$. The upper bound on $\Gamma_a(y, c) - \Gamma_a(y, c-1)$ reduces to $\min\{\Gamma_a(\infty, c) - \Gamma_a(\infty, c-1), \Gamma_a(y, \infty|s, a) - \max\{0, \Gamma_a(\infty, c-1|s, a)+\Gamma_a(y, \infty|s, a)-1\}\}$. Since $0 \geq \Gamma_a(\infty, c-1|s, a)+\Gamma_a(y, \infty|s, a)-1$, the upper bound on $\Gamma_a(y, c) - \Gamma_a(y, c-1)$ further reduces to $\min\{\Gamma_a(\infty, c) - \Gamma_a(\infty, c-1), \Gamma_a(y, \infty|s, a)\}$, which is the original upper bound given for $\Gamma_a^*(y, 0 \mid s, a)$. (4) $\Gamma_a(\infty, c|s, a) > \Gamma_a(y, \infty|s, a)$ and $0 < \Gamma_a(\infty, c-1|s, a)+\Gamma_a(y, \infty|s, a)-1$. The upper bound on $\Gamma_a(y, c)-\Gamma_a(y, c-1)$ further reduces to $\min\{\Gamma_a(\infty, c) - \Gamma_a(\infty, c-1), 1 - \Gamma_a(\infty, c-1|s, a)\} = \Gamma_a(\infty, c) - \Gamma_a(\infty, c-1)$. This implies that $\Gamma_a(\infty, c|s, a)-\Gamma_a(\infty, c-1|s, a) < \Gamma_a(y, \infty|s, a)$. The upper bound on $\Gamma_a^*(y, 0 \mid s, a)$ then also becomes $\Gamma_a(\infty, c|s, a) - \Gamma_a(\infty, c-1|s, a)$.

Finally, we show that the lower bound on $\Gamma_a(y, c) - \Gamma_a(y, c-1)$ is identical to the upper bound on $\Gamma_a^*(y, 0 \mid s, a) - \Gamma_a^*(y, -1 \mid s, a)$ in each of the following three possible cases. (1) $0 \geq \Gamma_a(\infty, c|s, a) + \Gamma_a(y, \infty|s, a) - 1$. The lower bound on $\Gamma_a(y, c) - \Gamma_a(y, c-1)$ reduces to $\max\{0, -\min\{\Gamma_a(\infty, c-1|s, a), \Gamma_a(y, \infty|s, a)\}\} = 0$. Because $\Gamma_a(\infty, c|s, a) \geq \Gamma_a(\infty, c|s, a) - \Gamma_a(\infty, c-1|s, a)$, the lower bound on $\Gamma_a^*(y, 0 \mid s, a)$ also becomes 0. (2) $0 < \Gamma_a(\infty, c|s, a) + \Gamma_a(y, \infty|s, a) - 1$ and $\Gamma_a(\infty, c-1|s, a) \leq \Gamma_a(y, \infty|s, a)$. The lower bound on $\Gamma_a(y, c) - \Gamma_a(y, c-1)$ reduces to $\max\{0, \Gamma_a(\infty, c|s, a) + \Gamma_a(y, \infty|s, a) - 1 - \min\{\Gamma_a(\infty, c-1|s, a), \Gamma_a(y, \infty|s, a)\}\}$. Since $\Gamma_a(\infty, c-1|s, a) \leq \Gamma_a(y, \infty|s, a)$,

the lower bound on $\Gamma_a(y,c)-\Gamma_a(y,c-1)$ reduces further to $\max\{0,\Gamma_a(\infty,c|s,a)+\Gamma_a(y,\infty|s,a)-1-\Gamma_a(\infty,c-1|s,a)\}$, which is the original lower bound given for $\Gamma_a^*(y,0 \mid s,a)$. (3) $0<\Gamma_a(\infty,c|s,a)+\Gamma_a(y,\infty|s,a)-1$ and $\Gamma_a(\infty,c-1|s,a)>\Gamma_a(y,\infty|s,a)$. The lower bound on $\Gamma_a(y,c)-\Gamma_a(y,c-1)$ reduces further to $\max\{0,\Gamma_a(\infty,c|s,a)+\Gamma_a(y,\infty|s,a)-1-\Gamma_a(y,\infty|s,a)\}=0$. Since $\Gamma_a(\infty,c|s,a)-\Gamma_a(\infty,c-1|s,a)+\Gamma_a(y,\infty|s,a)-1<\Gamma_a(\infty,c|s,a)-\Gamma_a(y,\infty|s,a)+\Gamma_a(y,\infty|s,a)-1<0$ and $\Gamma_a(\infty,c|s,a)-1\leq 0$, the lower bound for $\Gamma_a^*(y,0 \mid s,a)$ is also zero. ∎

Lemma .3 *Let $\Phi_a(y|s,c)=\Pr(Y_i(a)\leq y|S_i=s,C_i=c)$. Under Assumptions 1 and 2, the sharp upper and lower bounds on $\Phi_a(y|s,0)$, denoted by $\Phi_a^+(y|s,0)$ and $\Phi_a^-(y|s,0)$ respectively, are identified as*

$$\begin{aligned}\Phi_a^+(y|s,0) &= \min\left\{1,\ \frac{F(y|s,a,1)-F(y|s,a,0)P(a|s,0)}{P(0|s,0)}\right\} \text{ and}\\ \Phi_a^-(y|s,0) &= \max\left\{0,\ 1+\frac{P(a|s,0)+F(y|s,a,1)-F(y|s,a,0)P(a|s,0)-1}{P(0|s,0)}\right\}\end{aligned}$$

for $y\in\mathcal{Y}$ and $a,s\in\mathcal{A}$.

Proof. First, note that

$$\begin{aligned}\Phi_a(y|s,c) &= \Pr(Y_i(a)\leq y|S_i=s,C_i=c,C_i\neq a)\\ &= \frac{\Pr(Y_i(a)\leq y,C_i\leq c|S_i=s,C_i\neq a)-\Pr(Y_i(a)\leq y,C_i\leq c-1|S_i=s,C_i=\neq a)}{\Pr(C_i=c|S_i=s,C_i\neq a)}\\ &= \frac{\Gamma_a(y,c|s,a)-\Gamma_a(y,c-1|s,a)}{\Pr(C_i=c|S_i=s,C_i\neq a)},\end{aligned}$$

for $c\neq a$. By Lemma .1, the sharp upper and lower bounds on $\Phi_a(y|s,c)$ are given by

$$\begin{aligned}\Phi_a^+(y|s,c) &= \min\left\{1,\ \frac{\Gamma_a^+(y,c|s,a)-\Gamma_a^-(y,c-1|s,a)}{\Pr(C_i=c|S_i=s,C_i\neq a)}\right\},\\ \Phi_a^-(y|s,c) &= \max\left\{0,\ \frac{\Gamma_a^-(y,c|s,a)-\Gamma_a^+(y,c-1|s,a)}{\Pr(C_i=c|S_i=s,C_i\neq a)}\right\}.\end{aligned}$$

Because $\Gamma_a^+(y,-1|s,a)=\Gamma_a^-(y,-1|s,a)=0$ and by Lemma .1, these bounds simplify when $c=0$ to

$$\Phi_a^+(y|s,0) = \frac{\Gamma_a^+(y,0|s,a)}{\Pr(C_i=0|S_i=s,C_i\neq a)}$$

$$
\begin{aligned}
&= \min\left\{\frac{H(0|s,a,0)}{\Pr(C_i=0|S_i=s,C_i\neq a)},\ \frac{F(y|s,a,1)-F(y|s,a,0)P(a|s,0)}{\Pr(C_i=0|S_i=s,C_i\neq a)\left\{1-P(a|s,0)\right\}}\right\}\\
&= \min\left\{\frac{H(0|s,a,0)}{\Pr(A_i=0|S_i=s,A_i\neq a,D_i=0)},\right.\\
&\qquad\left.\frac{F(y|s,a,1)-F(y|s,a,0)P(a|s,0)}{\Pr(A_i=0|S_i=s,A_i\neq a,D_i=0)\Pr(A_i\neq a|S_i=s,D_i=0)}\right\}\\
&= \min\left\{1,\frac{F(y|s,a,1)-F(y|s,a,0)P(a|s,0)}{\Pr(A_i=0|S_i=s,D_i=0)}\right\}
\end{aligned}
$$

and

$$
\begin{aligned}
\Phi_a^-(y|s,0) &= \frac{\Gamma_a^-(y,0|s,a)}{\Pr(C_i=0|S_i=s,C_i\neq a)}\\
&= \max\left\{0,\ \frac{F(y|s,a,1)-F(y|s,a,0)P(a|s,0)-\left\{1-H(0|s,a,0)\right\}\left\{1-P(a|s,0)\right\}}{\Pr(C_i=0|S_i=s,C_i\neq a)\left\{1-P(a|s,0)\right\}}\right\}\\
&= \max\left\{0,\ \frac{F(y|s,a,1)-F(y|s,a,0)P(a|s,0)-1+P(a|s,0)}{\Pr(A_i=0|S_i=s,D_i=0)}+1\right\}. \qquad \blacksquare
\end{aligned}
$$

Now we provide a proof for the bounds in Proposition 1. We only consider the case of $c=0$. This can be done without loss of generality by Lemma .2. Now, note that $\tau(a,a'|0)$ can be written under Assumption 1 as,

$$
\tau(a,a'|0) \;=\; \sum_{s\in\mathcal{A}}\left\{\pi(a|s,0)-\pi(a'|s,0)\right\}\Pr(S_i=s|A_i=0,D_i=0), \tag{13}
$$

where $\pi(a|s,c)\equiv\mathbb{E}[Y_i(a)|S_i=s,C_i=c]$ for any a and $c\in\mathcal{A}$. Under Assumption 1, $\pi(a|s,0)$ can be point-identified when $a=0$ as

$$
\pi(0|s,0) \;=\; \mathbb{E}[Y_i|A_i=0,S_i=s,D_i=0], \tag{14}
$$

for any $s\in\mathcal{A}$, but not when $a\neq 0$. To find the sharp bounds on $\pi(a|s,0)$ when $a\neq 0$, note that

$$
\pi(a|s,0) \;=\; \lim_{y^*\to-\infty}\left\{\int_{y^*}^{\infty}1-\Phi_a(y|s,0)\,\mathrm{d}y+y^*\right\}.
$$

By Lemma .3, $\pi^-(a|s,0)\leq\pi(a|s,0)\leq\pi^+(a|s,0)$ where

$$
\pi^-(a|s,0) \;\equiv\; \lim_{y^*\to-\infty}\left\{\int_{y^*}^{\infty}1-\Phi_a^+(y|s,0)\,\mathrm{d}y+y^*\right\}, \tag{15}
$$

$$
\pi^+(a|s,0) \;\equiv\; \lim_{y^*\to-\infty}\left\{\int_{y^*}^{\infty}1-\Phi_a^-(y|s,0)\,\mathrm{d}y+y^*\right\}. \tag{16}
$$

The bounds, $\pi^-(a|s,0)$ and $\pi^+(a|s,0)$, are the sharp lower and upper bounds on $\pi(a|s,0)$ because $\Phi_a^+(y|s,0)$ and $\Phi_a^-(y|s,0)$ are the sharp upper and lower bounds on $\Phi_a(y|s,0)$, respectively.

Substituting Equations (14), (15) and (16) into Equation (13) and simplifying the terms yield the sharp bounds on $\tau(a,0|0)$,

$$\begin{aligned}\sum_{s\in\mathcal{A}}\left\{\pi^-(a|s,0)\Pr(S_i=s|A_i=0,D_i=0)\right\}-\mathbb{E}[Y_i|A_i=0,D_i=0] \\ \leq\ \tau(a,0|0)\ \leq \\ \sum_{s\in\mathcal{A}}\left\{\pi^+(a|s,0)\Pr(S_i=s|A_i=0,D_i=0)\right\}-\mathbb{E}[Y_i|A_i=0,D_i=0]\end{aligned} \tag{17}$$

for any $a\in\mathcal{A}$. For $\tau(a,a')$ where $a\neq a'$, we obtain the following bounds,

$$\begin{aligned}\sum_{s\in\mathcal{A}}\left\{\pi^-(a|s,0)-\pi^+(a'|s,0)\right\}\Pr(S_i=s|A_i=0,D_i=0) \\ \leq\ \tau(a,a'|0)\ \leq \\ \sum_{s\in\mathcal{A}}\left\{\pi^+(a|s,0)-\pi^-(a'|s,0)\right\}\Pr(S_i=s|A_i=0,D_i=0)\end{aligned} \tag{18}$$

which are not necessarily sharp because $\pi^-(a|s,0)$ and $\pi^+(a'|s,0)$ may not be simultaneously attainable, and vice versa. Finally, Lemma .2 implies that (17) and (18) are both valid as bounds for $\tau(a,c|c)$ and $\tau(a,a'|c)$, respectively, for any c $\in\mathcal{A}$. This completes the proof of Proposition 1. ∎

A.4 Proof of Proposition 2

We begin by considering the joint distribution of all variables in the study population when $J=3$:

$$\begin{aligned}&\Pr(S_i=s,D_i=d,C_i=c,A_i=a,Y_i=y,Y_i(0)=y_0,Y_i(1)=y_1,Y_i(2)=y_2) \\ &=\ \Pr(Y_i(d)=y|A_i=a,Y_i(0)=y_0,Y_i(1)=y_1,Y_i(2)=y_2) \\ &\quad\times\Pr(A_i=a|C_i=c,D_i=d) \\ &\quad\times\Pr(S_i=s,C_i=c,Y_i(0)=y_0,Y_i(1)=y_1,Y_i(2)=y_2)\Pr(D_i=d) \\ &=\ \Pr(Y_i(d)=y|A_i=a,Y_i(0)=y_0,Y_i(1)=y_1,Y_i(2)=y_2) \\ &\quad\times\{\Pr(A_i=a|C_i=c,D_i=0)(1-d)+\Pr(A_i=a|D_i=1)d\}\end{aligned}$$

$$\times \Pr(S_i = s, C_i = c, Y_i(0) = y_0, Y_i(1) = y_1, Y_i(2) = y_2) \Pr(D_i = d), \tag{19}$$

where the first equality follows from Assumption 1 and the fact that $Y_i(0)$, $Y_i(1)$, $Y_i(2)$ and A_i are sufficient for Y_i and that C_i and D_i are sufficient for A_i. The second equality is by Assumption 2. Note that $\Pr(Y_i(d) = y|Y_i(0), Y_i(1), Y_i(2))$ and $\Pr(A_i = a|C_i, D_i = 0)$ are degenerate and that $\Pr(A_i = a|D_i = 1)$ and $\Pr(D_i = d)$ are fixed by the experimental design. Therefore, the remaining component of equation (19), $\Pr(S_i = s, C_i = c, Y_i(0) = y_0, Y_i(1) = y_1, Y_i(2) = y_2)$, completely specifies the data generating process, with $|\mathcal{A}|^2 \cdot |\mathcal{Y}|^{|\mathcal{A}|} - 1 = J^2 2^J - 1$ free parameters needed to describe it. Balke (1995, Section 3.5) shows that bounds on counterfactual probabilities found by optimizing over such a complete model are sharp; that is, they are guaranteed to be at least as tight as bounds calculated from any partial (marginalized) model.

We express the complete model in terms of $\phi_{y_0,y_1,y_2,s,c} \in \Phi$. First, note that $\sum_{y_0\in\{0,1\}} \sum_{y_1\in\{0,1\}} \sum_{y_2\in\{0,1\}} \sum_{s'\in\mathcal{A}} \sum_{c'\in\mathcal{A}} \phi_{y_a,y_{a'},y_{a''},s',c'} = 1$. Next, from the free-choice condition, we observe $\Pr(S_i = s, C_i = c, Y_i = y \mid D_i = 0)$, which is completely specified by $|\mathcal{A}|^2 \cdot |\mathcal{Y}| - 1 = 2J^2 - 1$ free parameters. We use the following $2J^2$ marginals as constraints on $\phi_{y_0,y_1,y_2,s,c}$ (with one redundant):

$$\Pr(S_i = s, A_i = c \mid D_i = 0) = \Pr(S_i = s, C_i = c) = \sum_{a\in\mathcal{A}} \sum_{y_a\in\{0,1\}} \phi_{y_0,y_1,y_2,s,c}, \tag{20}$$

$$\Pr(S_i = s, A_i = c, Y_i = 1 \mid D_i = 0) = \Pr(S_i = s, C_i = c, Y_i(c) = 1) = \sum_{a\neq c} \sum_{y_a\in\{0,1\}} \phi_{y_0,y_1,y_2,s,c}, \tag{21}$$

for all s and $c \in \mathcal{A}$. Similarly, from the forced-choice condition, we observe

$$\begin{aligned}&\Pr(S_i = s, A_i = a, Y_i = y \mid D_i = 1)\\ &\quad = \Pr(Y_i = y \mid S_i = s, A_i = a, D_i = 1) \Pr(A_i = a \mid D_i = 1) \Pr(S_i = s \mid D_i = 1)\end{aligned}$$

where the equality holds by Assumption 2. Because $\Pr(A_i = a \mid D_i = 1)$ is fixed a priori by randomization, the observed distribution from the forced-choice arm can be fully characterized by $(|\mathcal{Y}| - 1)|\mathcal{A}|^2 + |\mathcal{A}| - 1 = J^2 + J - 1$ free parameters. We use the following $J^2 + J$ margins as constraints on $\phi_{y_0,y_1,y_2,s,c}$,

noting that one of them is redundant:

$$\Pr(S_i = s \mid A_i = a, D_i = 1) \;=\; \Pr(S_i = s) \;=\; \sum_{a\in\mathcal{A}} \sum_{y_a\in\{0,1\}} \sum_{c\in\mathcal{A}} \phi_{y_0,y_1,y_2,s,c}, \tag{22}$$

$$\Pr(S_i = s, Y_i = 1 \mid A_i = a, D_i = 1) \;=\; \Pr(S_i = s, Y_i(a) = 1) \;=\; \sum_{a'\in\mathcal{A}} \sum_{y_{a'}\in\{0,1\}} \sum_{c\in\mathcal{A}} \phi_{y_0,y_1,y_2,s,c} \cdot \mathbf{1}\{y_a = 1\},$$

for all s and $a \in \mathcal{A}$. However, note that equation (22) are merely linear combinations of equation (20) and can therefore be omitted.

Finally, the quantity of interest can be written in terms of $\phi_{y_0,y_1,y_2,s,c}$ as,

$$\begin{aligned}\tau(a, a' \mid c) \;&=\; \mathbb{E}[Y_i(a) \mid C_i = c] - \mathbb{E}[Y_i(a') \mid C_i = c] \\ &=\; \frac{\sum_{y_0\in\{0,1\}} \sum_{y_2\in\{0,1\}} \sum_s \phi_{1,y_1,y_2,s,c}}{\Pr(A_i = c|D_i = 0)} \;-\; \frac{\sum_{y_1\in\{0,1\}} \sum_{y_2\in\{0,1\}} \sum_s \phi_{y_0,1,y_2,s,c}}{\Pr(A_i = c|D_i = 0)},\end{aligned}$$

assuming $a' = 1$ and $a = 0$ without loss of generality. Solving for the extrema of $\tau(a, a' \mid c)$ under the above set of linear constraints, which incorporate the full information in the observed data as well as probability axioms, yields its sharp upper and lower bounds as displayed in Proposition 2. ∎

A.5 Statistical Inference for the Bounds

Let $\boldsymbol{p} = [p_s] = [\Pr(S_i = 0), \cdots, \Pr(S_i = J-1)]^\top$ be a stochastic vector of stated-preference probabilities. $\boldsymbol{q} = [q_{sc}] = [\Pr(C_i = c|S_i = s)]$ is a row-stochastic matrix, where row s, denoted $\boldsymbol{q}_s$, represents the distribution of true preferences (C_i) among those with the stated preference $S_i = s$. Also let $\boldsymbol{\pi}^+ = \{\pi^+(a|s,c) : a, s, c \in \mathcal{A}\}$ and $\boldsymbol{\pi}^- = \{\pi^-(a|s,c) : a, s, c \in \mathcal{A}\}$, where $\pi^+(a|s,c)$ and $\pi^-(a|s,c)$ are defined in Appendix A.3. Let $\boldsymbol{F}^1 = \{F(y|s,a,d) : s, a \in \mathcal{A}, d = 1\}$ and $\boldsymbol{F}^0 = \{F(y|s,a,d) : s, a \in \mathcal{A}, d = 0\}$, where $F(y|s,a,d)$ is defined in Proposition 1. Finally, we use $\boldsymbol{\tau}^+$ and $\boldsymbol{\tau}^-$ to denote the sets of the upper and lower bounds on $\tau(a, a'|c)$ for all $a, a', c \in \mathcal{A}$, respectively, and $\boldsymbol{X}$ to indicate all observed data.

Our goal is to approximate the posterior distribution of $(\boldsymbol{\tau}^-, \boldsymbol{\tau}^+)$ with Monte Carlo simulations. We begin by the general bounds in Proposition 1. Note that $\boldsymbol{\tau}^-$ and $\boldsymbol{\tau}^+$ are deterministic functions of $\boldsymbol{\pi}^-$,

$\boldsymbol{\pi}^+$, $\boldsymbol{p}$ and $\boldsymbol{q}$, such that

$$\begin{aligned}\tau^-(a,a'|c) &= \sum_{s\in\mathcal{A}}\left(\pi^-(a|s,c)-\pi^+(a'|s,c)\right)\frac{q_{sc}p_s}{\sum_{s'\in\mathcal{A}}q_{s'c}p_{s'}},\\ \tau^+(a,a'|c) &= \sum_{s\in\mathcal{A}}\left(\pi^+(a|s,c)-\pi^-(a'|s,c)\right)\frac{q_{sc}p_s}{\sum_{s'\in\mathcal{A}}q_{s'c}p_{s'}}\end{aligned}$$

for all $a, a', c \in \mathcal{A}$. Therefore, we consider the problem of simulating samples from the joint posterior of $\boldsymbol{\pi}^-$, $\boldsymbol{\pi}^+$, $\boldsymbol{p}$ and $\boldsymbol{q}$, which can be written as,

$$f(\boldsymbol{\pi}^+,\boldsymbol{\pi}^-,\boldsymbol{p},\boldsymbol{q}|\boldsymbol{X}) = f(\boldsymbol{\pi}^+,\boldsymbol{\pi}^-|\hat{\boldsymbol{F}}^1,\hat{\boldsymbol{F}}^0,\boldsymbol{q})\; f(\boldsymbol{q}|\boldsymbol{n}_s^0)\; f(\boldsymbol{p}|\boldsymbol{n})$$

under Assumptions 1 and 2, where $\hat{\boldsymbol{F}}^1$ and $\hat{\boldsymbol{F}}^0$ are empirical CDFs corresponding to $\boldsymbol{F}^1$ and $\boldsymbol{F}^0$, respectively. For $\boldsymbol{p}$ and $\boldsymbol{q}$, we use the noninformative improper priors $\boldsymbol{p} \sim$ Dirichlet$(\boldsymbol{0})$ and $\boldsymbol{q}_s \sim$ Dirichlet$(\boldsymbol{0})$ $\forall\, s \in \mathcal{A}$. Then, $\boldsymbol{q}_s \mid \boldsymbol{X} \sim$ Dirichlet$(\boldsymbol{n}_s^0)$ $\forall\, s$ and $\boldsymbol{p} \mid \boldsymbol{X} \sim$ Dirichlet$(\boldsymbol{n})$.

We are now left with $f(\boldsymbol{\pi}^+,\boldsymbol{\pi}^-|\hat{\boldsymbol{F}}^1,\hat{\boldsymbol{F}}^0,\boldsymbol{q})$. Because of the way these bounds are constructed (see Proposition 1),

$$\begin{aligned}&\pi^+(a|s,c),\pi^-(a|s,c) \perp\!\!\!\perp \pi^+(a|s',c),\pi^-(a|s',c) \mid \hat{\boldsymbol{F}}^1,\hat{\boldsymbol{F}}^0,\boldsymbol{q} \quad \text{and}\\ &\pi^+(a|s,c),\pi^-(a|s,c) \perp\!\!\!\perp \pi^+(a'|s,c),\pi^-(a'|s,c) \mid \hat{\boldsymbol{F}}^1,\hat{\boldsymbol{F}}^0,\boldsymbol{q}\end{aligned}$$

for $s \neq s'$ and $a \neq a'$. Therefore, to fully characterize the posterior of $[\tau^-(a',a''|c),\tau^+(a',a''|c)]$ for each a, a'' and $c \in \mathcal{A}$, it is sufficient to only consider the bivariate posterior distribution of $[\pi^+(a|s,c),\pi^-(a|s,c)]$ for $a \in \{a',a''\}$ and $s \in \mathcal{A}$. Note that, under mild assumptions and with a sufficiently large sample size, the posterior for each pair $[\pi^+(a|s,c),\pi^-(a|s,c)]$ can be approximated by a bivariate normal distribution due to the Bayesian central limit theorem. That is, we have:

$$\begin{bmatrix}\pi^-(a|s,c)\\ \pi^+(a|s,c)\end{bmatrix} \mid \boldsymbol{q},\boldsymbol{X} \;\approx\; \text{Normal}\left(\begin{bmatrix}\bar{\pi}^-(a|s,c,\boldsymbol{q}_s,\boldsymbol{X})\\ \bar{\pi}^+(a|s,c,\boldsymbol{q}_s,\boldsymbol{X})\end{bmatrix},\begin{bmatrix}V^-(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) & C(a|s,c,\boldsymbol{q}_s,\boldsymbol{X})\\ C(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) & V^+(a|s,c,\boldsymbol{q}_s,\boldsymbol{X})\end{bmatrix}\right) \tag{23}$$

when N is sufficiently large, and the means and covariances can be approximated by the asymptotic means and covariances of the frequentist sampling distributions of $[\pi^-(a|s,c),\pi^+(a|s,c)]$, respectively, as shown below. Note that priors on $\pi^-(a|s,c)$, $\pi^+(a|s,c)$ can be ignored and therefore left unspecified

when N is large because of the Bernstein-von Mises theorem.

Let $\underline{y}$ be the natural lower bound of $Y_i(a)$ if it exists and $\min\{Y_i : S_i = s, A_i = a\}$, which is the lowest point at which the estimated conditional CDF, $\hat{\Gamma}_a(y, \infty|s, a)$, is nonzero, if it does not. Let $\Gamma_a^{-1}(\cdot)$ be the inverse of $\Gamma_a(y, \infty|s, a)$ (see Section A.3 for the definition) with respect to y, so that $\Gamma_a^{-1}(\Gamma_a(y, \infty|s, a)) = y$, and let $\hat{\Gamma}_a^{-1}(\cdot)$ be its sample analogue, such that $\hat{\Gamma}_a^{-1}(p) = \min\{y : p \leq \hat{\Gamma}_a(y, \infty|s, a)\}$. Let $b = \frac{q_{sc}}{1-q_{sa}}$. For the means, note that the $\pi^-(a|s, c)$ and $\pi^+(a|s, c)$ are functions of $F(y|s, a, 0)$, $F(y|s, a, 1)$ and $P(a|s, 0)$ (as shown in Appendix A.3), which can be consistently estimated by their nonparametric maximum likelihood estimates $\hat{F}(y|s, a, 0)$, $\hat{F}(y|s, a, 1)$ and q_{sa}, respectively. This implies the following plug-in estimators for $\bar{\pi}^-(a|s, c, \boldsymbol{q}_s, \boldsymbol{X})$ and $\bar{\pi}^+(a|s, c, \boldsymbol{q}_s, \boldsymbol{X})$:

$$\begin{aligned}
\hat{\bar{\pi}}^-(a|s, c, \boldsymbol{q}_s, \boldsymbol{X}) &= \hat{\Gamma}_a^{-1}(b) - \int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(b)} \frac{\hat{F}(y|s, a, 1) - \hat{F}(y|s, a, 0)q_{sa}}{q_{sc}} \, \mathrm{d}y \\
\hat{\bar{\pi}}^+(a|s, c, \boldsymbol{q}_s, \boldsymbol{X}) &= \hat{\Gamma}_a^{-1}(1-b) - \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\infty} \frac{q_{sa} + \hat{F}(y|s, a, 1) - \hat{F}(y|s, a, 0)q_{sa} - 1}{q_{sc}} \, \mathrm{d}y,
\end{aligned}$$

where we used the fact that $\Phi_a^+(y|s, c) = 1$ for $y \geq \Gamma_a^{-1}(b)$ and $\Phi_a^-(y|s, c) = 0$ for $y \leq \Gamma_a^{-1}(1-b)$ (see Appendix A.3 for the definitions of $\Phi_a^+(y|s, c)$ and $\Phi_a^-(y|s, c)$).

For the variances and covariances, we use the fact that for any ECDF $\hat{F}(\cdot)$, $\mathrm{Cov}\left(\hat{F}(a), \hat{F}(b)\right) = \frac{F(a)-F(a)F(b)}{n}$ for $a \leq b$ where n is the number of steps in $\hat{F}(\cdot)$.

$$\begin{aligned}
&V^-(a|s, c, \boldsymbol{q}_s, \boldsymbol{X}) \\
&\quad= \mathrm{Var}\left(\hat{\Gamma}_a^{-1}(b) - \int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(b)} \frac{\hat{F}(y|s, a, 1) - \hat{F}(y|s, a, 0)q_{sa}}{q_{sc}} \, \mathrm{d}y\right) \\
&\quad= \left(\frac{1}{q_{sc}}\right)^2 \mathrm{Var}\left(\int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(b)} \hat{F}(y|s, a, 1) - \hat{F}(y|s, a, 0)q_{sa} \, \mathrm{d}y\right) \\
&\quad= \left(\frac{1}{q_{sc}}\right)^2 \int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(b)} \int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(b)} \mathrm{Cov}\begin{pmatrix} \hat{F}(y|s, a, 1) - \hat{F}(y|s, a, 0)q_{sa}, \\ \hat{F}(x|s, a, 1) - \hat{F}(x|s, a, 0)q_{sa} \end{pmatrix} \mathrm{d}x\mathrm{d}y \\
&\quad= 2\left(\frac{1}{q_{sc}}\right)^2 \int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(b)} \int_{y}^{\hat{\Gamma}_a^{-1}(b)} \mathrm{Cov}\begin{pmatrix} \hat{F}(y|s, a, 1) - \hat{F}(y|s, a, 0)q_{sa}, \\ \hat{F}(x|s, a, 1) - \hat{F}(x|s, a, 0)q_{sa} \end{pmatrix} \mathrm{d}x\mathrm{d}y
\end{aligned}$$

$$
\begin{aligned}
&= 2\left(\frac{1}{q_{sc}}\right)^2 \int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(b)} \int_{y}^{\hat{\Gamma}_a^{-1}(b)} \mathrm{Cov}\left(\hat{F}(y|s,a,1), \hat{F}(x|s,a,1)\right) \,\mathrm{d}x\mathrm{d}y \\
&\quad + 2\left(\frac{q_{sa}}{q_{sc}}\right)^2 \int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(b)} \int_{y}^{\hat{\Gamma}_a^{-1}(b)} \mathrm{Cov}\left(\hat{F}(y|s,a,0), \hat{F}(x|s,a,0)\right) \,\mathrm{d}x\mathrm{d}y \\
&= \frac{2}{n_{sa}^1}\left(\frac{1}{q_{sc}}\right)^2 \int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(b)} \int_{y}^{\hat{\Gamma}_a^{-1}(b)} F(y|s,a,1)\left(1 - F(x|s,a,1)\right) \,\mathrm{d}x\mathrm{d}y \\
&\quad + \frac{2}{n_{sa}^0}\left(\frac{q_{sa}}{q_{sc}}\right)^2 \int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(b)} \int_{y}^{\hat{\Gamma}_a^{-1}(b)} F(y|s,a,0)\left(1 - F(x|s,a,0)\right) \,\mathrm{d}x\mathrm{d}y,
\end{aligned}
$$

where n_{sa}^0 is as defined in Section 6 and $n_{sa}^1 = \sum_{i=1}^N \mathbf{1}\{S_i = s, A_i = a, D_i = 1\}$. Similarly,

$$
\begin{aligned}
&V^+(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) \\
&\quad = \mathrm{Var}\left(\hat{\Gamma}_a^{-1}(1-b) - \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\infty} \frac{q_{sa} + \hat{F}(y|s,a,1) - \hat{F}(y|s,a,0)q_{sa} - 1}{q_{sc}} \,\mathrm{d}y\right) \\
&\quad = \frac{2}{n_{sa}^1}\left(\frac{1}{q_{sc}}\right)^2 \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\infty} \int_{y}^{\infty} F(y|s,a,1)\left(1 - F(x|s,a,1)\right) \,\mathrm{d}x\mathrm{d}y \\
&\qquad + \frac{2}{n_{sa}^0}\left(\frac{q_{sa}}{q_{sc}}\right)^2 \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\infty} \int_{y}^{\infty} F(y|s,a,0)\left(1 - F(x|s,a,0)\right) \,\mathrm{d}x\mathrm{d}y
\end{aligned}
$$

We estimate these quantities by substituting $F(\cdot|s,a,d)$ with $\hat{F}(\cdot|s,a,d)$ for $d = 0, 1$. A small sample correction can optionally be applied to these estimates by replacing n_{sa}^d with $n_{sa}^d - 1$ for $d = 0, 1$.

The covariance between $\pi^-(a|s,c)$ and $\pi^+(a|s,c)$ depends on whether $b < \frac{1}{2}$, in which case they are based on disjoint (but still correlated) portions of the same ECDFs, or whether $b \geq \frac{1}{2}$, in which case they are based on overlapping regions of the ECDFs and are therefore more correlated. If $b \geq \frac{1}{2}$,

$$
\begin{aligned}
&C\left(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}\right) \\
&\quad = \mathrm{Cov}\Bigg(\hat{\Gamma}_a^{-1}(b) - \int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(b)} \frac{\hat{F}(y|s,a,1) - \hat{F}(y|s,a,0)q_{sa}}{q_{sc}} \,\mathrm{d}y, \\
&\qquad\qquad \hat{\Gamma}_a^{-1}(1-b) - \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\infty} \frac{q_{sa} + \hat{F}(y|s,a,1) - \hat{F}(y|s,a,0)q_{sa} - 1}{q_{sc}} \,\mathrm{d}y\Bigg) \\
&\quad = \mathrm{Cov}\left(\int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(b)} \frac{\hat{F}(y|s,a,1) - \hat{F}(y|s,a,0)q_{sa}}{q_{sc}} \,\mathrm{d}y, \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\infty} \frac{\hat{F}(y|s,a,1) - \hat{F}(y|s,a,0)q_{sa}}{q_{sc}} \,\mathrm{d}y\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\frac{1}{q_{sc}}\right)^2 \int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(1-b)} \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\infty} \mathrm{Cov}\left(\begin{array}{c} \hat{F}(y|s,a,1) - \hat{F}(y|s,a,0)q_{sa}, \\ \hat{F}(x|s,a,1) - \hat{F}(x|s,a,0)q_{sa} \end{array}\right) \mathrm{d}x\mathrm{d}y \\
&\quad + 2\left(\frac{1}{q_{sc}}\right)^2 \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\hat{\Gamma}_a^{-1}(b)} \int_{y}^{\hat{\Gamma}_a^{-1}(b)} \mathrm{Cov}\left(\begin{array}{c} \hat{F}(y|s,a,1) - \hat{F}(y|s,a,0)q_{sa}, \\ \hat{F}(x|s,a,1) - \hat{F}(x|s,a,0)q_{sa} \end{array}\right) \mathrm{d}x\mathrm{d}y \\
&\quad + \left(\frac{1}{q_{sc}}\right)^2 \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\hat{\Gamma}_a^{-1}(b)} \int_{\hat{\Gamma}_a^{-1}(b)}^{\infty} \mathrm{Cov}\left(\begin{array}{c} \hat{F}(y|s,a,1) - \hat{F}(y|s,a,0)q_{sa}, \\ \hat{F}(x|s,a,1) - \hat{F}(x|s,a,0)q_{sa} \end{array}\right) \mathrm{d}x\mathrm{d}y \\
&= \frac{1}{n_{sa}^1}\left(\frac{1}{q_{sc}}\right)^2 \int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(1-b)} \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\infty} F(y|s,a,1)\left(1 - F(x|s,a,1)\right)\, \mathrm{d}x\mathrm{d}y \\
&\quad + \frac{1}{n_{sa}^0}\left(\frac{q_{sa}}{q_{sc}}\right)^2 \int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(1-b)} \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\infty} F(y|s,a,0)\left(1 - F(x|s,a,0)\right)\, \mathrm{d}x\mathrm{d}y \\
&\quad + \frac{2}{n_{sa}^1}\left(\frac{1}{q_{sc}}\right)^2 \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\hat{\Gamma}_a^{-1}(b)} \int_{y}^{\hat{\Gamma}_a^{-1}(b)} F(y|s,a,1)\left(1 - F(x|s,a,1)\right)\, \mathrm{d}x\mathrm{d}y \\
&\quad + \frac{2}{n_{sa}^0}\left(\frac{q_{sa}}{q_{sc}}\right)^2 \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\hat{\Gamma}_a^{-1}(b)} \int_{y}^{\hat{\Gamma}_a^{-1}(b)} F(y|s,a,0)\left(1 - F(x|s,a,0)\right)\, \mathrm{d}x\mathrm{d}y \\
&\quad + \frac{1}{n_{sa}^1}\left(\frac{1}{q_{sc}}\right)^2 \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\hat{\Gamma}_a^{-1}(b)} \int_{\hat{\Gamma}_a^{-1}(b)}^{\infty} F(y|s,a,1)\left(1 - F(x|s,a,1)\right)\, \mathrm{d}x\mathrm{d}y \\
&\quad + \frac{1}{n_{sa}^0}\left(\frac{q_{sa}}{q_{sc}}\right)^2 \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\hat{\Gamma}_a^{-1}(b)} \int_{\hat{\Gamma}_a^{-1}(b)}^{\infty} F(y|s,a,0)\left(1 - F(x|s,a,0)\right)\, \mathrm{d}x\mathrm{d}y
\end{aligned}
$$

and if $b < \frac{1}{2}$,

$$
\begin{aligned}
&C\left(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}\right) \\
&\quad = \frac{1}{n_{sa}^1}\left(\frac{1}{q_{sc}}\right)^2 \int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(b)} \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\infty} F(y|s,a,1)\left(1 - F(x|s,a,1)\right)\, \mathrm{d}x\mathrm{d}y \\
&\qquad + \frac{1}{n_{sa}^0}\left(\frac{q_{sa}}{q_{sc}}\right)^2 \int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(b)} \int_{\hat{\Gamma}_a^{-1}(1-b)}^{\infty} F(y|s,a,0)\left(1 - F(x|s,a,0)\right)\, \mathrm{d}x\mathrm{d}y.
\end{aligned}
$$

Again, we estimate these by replacing $F(\cdot|s,a,d)$ with $\hat{F}(\cdot|s,a,d)$ for $d = 0, 1$. The small sample correction can also be applied.

Finally, in the special case of $a = c$, the quantity $\pi(a|s,c) = \pi(c|s,c)$ is point-identified. Therefore, equation (23) reduces to a univariate normal distribution such that $\bar{\pi} \equiv \bar{\pi}^-(c|s,c,\boldsymbol{q}_s,\boldsymbol{X}) = \bar{\pi}^+(c|s,c,\boldsymbol{q}_s,\boldsymbol{X})$ and $V \equiv V^-(c|s,c,\boldsymbol{q}_s,\boldsymbol{X}) = V^+(c|s,c,\boldsymbol{q}_s,\boldsymbol{X}) = C(c|s,c,\boldsymbol{q}_s,\boldsymbol{X})$. In fact, the es-

timators of these parameters provided above reduce to the sample mean and the sampling variance for the mean, respectively, for the corresponding subgroup:

$$\begin{aligned}
\hat{\bar{\pi}} &= \underline{y} + \int_{\underline{y}}^{\infty} 1 - \hat{F}(y|s,c,0)\mathrm{d}y \\
&= \underline{y} + \int_{\underline{y}}^{\infty} \sum_{i=1}^{N} \Big(1 - \mathbf{1}\{Y_i \le y\}\Big) \cdot \frac{\mathbf{1}\{S_i = s, A_i = c, D_i = 0\}}{n_{sc}^0}\,\mathrm{d}y \\
&= \underline{y} + \frac{1}{n_{sc}^0} \sum_{i=1}^{N} \left(\int_{\underline{y}}^{Y_i} 1\,\mathrm{d}y + \int_{Y_i}^{\infty} 0\,\mathrm{d}y \right) \cdot \mathbf{1}\{S_i = s, A_i = c, D_i = 0\} \\
&= \frac{1}{n_{sc}^0} \sum_{i=1}^{N} Y_i \cdot \mathbf{1}\{S_i = s, A_i = c, D_i = 0\},
\end{aligned}$$

and

$$\begin{aligned}
\hat{V} &= \frac{2}{n_{sc}^0} \int_{\underline{y}}^{\infty} \int_{y}^{\infty} \hat{F}(y|s,c,0) \Big(1 - \hat{F}(x|s,c,0)\Big)\,\mathrm{d}x\mathrm{d}y \\
&= \frac{2}{n_{sc}^0} \int_{\underline{y}}^{\infty} \int_{y}^{\infty} \left(\sum_{i=1}^{N} \mathbf{1}\{Y_i \le y\} \cdot \frac{\mathbf{1}\{S_i = s, A_i = c, D_i = 0\}}{n_{sc}^0} \right) \\
&\qquad \times \left(\sum_{j=1}^{N} \Big(1 - \mathbf{1}\{Y_j \le x\}\Big) \cdot \frac{\mathbf{1}\{S_j = s, A_j = c, D_j = 0\}}{n_{sc}^0} \right) \mathrm{d}x\mathrm{d}y \\
&= \frac{2}{(n_{sc}^0)^3} \int_{\underline{y}}^{\infty} \left(\sum_{i=1}^{N} \mathbf{1}\{Y_i \le y\} \cdot \mathbf{1}\{S_i = s, A_i = c, D_i = 0\} \right) \\
&\qquad \times \sum_{j=1}^{N} \left(\int_{y}^{\infty} \Big(1 - \mathbf{1}\{Y_j \le x\}\Big) \cdot \mathbf{1}\{S_j = s, A_j = c, D_j = 0\}\,\mathrm{d}x \right) \mathrm{d}y \\
&= \frac{2}{(n_{sc}^0)^3} \sum_{i=1}^{N} \sum_{j=1}^{N} \int_{\underline{y}}^{\infty} \mathbf{1}\{Y_i \le y\} \cdot \mathbf{1}\{S_i = s, A_i = c, D_i = 0\} \\
&\qquad \times \Big(1 - \mathbf{1}\{Y_j \le y\}\Big)(Y_j - y) \cdot \mathbf{1}\{S_j = s, A_j = c, D_j = 0\}\,\mathrm{d}y \\
&= \frac{2}{(n_{sc}^0)^3} \sum_{i=1}^{N} \sum_{j \in \mathcal{J}} \mathbf{1}\{S_i = s, A_i = c, D_i = 0\} \cdot \mathbf{1}\{S_j = s, A_j = c, D_j = 0\} \int_{Y_i}^{Y_j} (Y_j - y)\,\mathrm{d}y, \\
&\qquad \text{with } \mathcal{J} = \big\{j \in 1, \cdots, N : Y_j \ge Y_i\big\} \\
&= \frac{1}{n_{sc}^0} \sum_{i=1}^{N} \sum_{j \in \mathcal{J}} \frac{(Y_j - Y_i)^2}{(n_{sc}^0)^2} \cdot \mathbf{1}\{S_i = s, A_i = c, D_i = 0\} \cdot \mathbf{1}\{S_j = s, A_j = c, D_j = 0\} \\
&= \frac{1}{(n_{sc}^0)^2} \sum_{i=1}^{N} (Y_i - \bar{\pi})^2 \cdot \mathbf{1}\{S_i = s, A_i = c, D_i = 0\},
\end{aligned}$$

for any $c, s \in \mathcal{A}$. Again, a small sample correction can be applied for $\hat{V}$ by multiplying it by $n^0_{sc}/(n^0_{sc}-1)$.

For the binary-outcome bounds in Proposition 2, we employ a similar procedure. Let $\boldsymbol{H} = [H_{sa}] = [\Pr(Y_i = 1|S_i = s, A_i = a, D_i = 1)]$ and $\boldsymbol{G} = [G_{sa}] = [\Pr(Y_i = 1|S_i = s, A_i = a, D_i = 0)]$. In this case, $\boldsymbol{\tau}^-$ and $\boldsymbol{\tau}^+$ are completely determined by $\boldsymbol{H}$, $\boldsymbol{G}$, $\boldsymbol{p}$ and $\boldsymbol{q}$. The endpoints of the ACTE bounds $\tau^-(a, a'|c)$ and $\tau^+(a, a'|c)$ are respectively given by the solutions to the linear problem described in Proposition 2:

$$\min_{\Phi} \quad \text{and} \quad \max_{\Phi} \quad \frac{1}{\Pr(A_i = c|D_i = 0)} \left\{ \sum_{a'' \in \{0,1\}} \sum_{s \in \mathcal{A}} \left(\phi_{1,0,y_{a''},s,c} - \phi_{0,1,y_{a''},s,c} \right) \right\}, \tag{24}$$

s.t. $\phi_{y_0,y_1,y_2,s,c'} \geq 0 \ \forall \ y_0, y_1, y_2, s, c'$, $\sum_{y_0 \in \{0,1\}} \sum_{y_1 \in \{0,1\}} \sum_{y_2 \in \{0,1\}} \sum_{s \in \mathcal{A}} \sum_{c' \in \mathcal{A}} \phi_{y_0,y_1,y_2,s,c'} = 1$, $\sum_{y_0 \in \{0,1\}} \sum_{y_1 \in \{0,1\}} \sum_{y_2 \in \{0,1\}} \phi_{y_0,y_1,y_2,s,c'} \cdot \mathbf{1}\{y_{c'} = 1\} = q_{sc'} p_s G_{sc'} \ \forall \ s, c'$, $\sum_{y_0 \in \{0,1\}} \sum_{y_1 \in \{0,1\}} \sum_{y_2 \in \{0,1\}} \phi_{y_0,y_1,y_2,s,c'} = q_{sc'} p_s \ \forall \ s, c'$, and $\sum_{y_0 \in \{0,1\}} \sum_{y_1 \in \{0,1\}} \sum_{y_2 \in \{0,1\}} \sum_{c' \in \mathcal{A}} \phi_{y_0,y_1,y_2,s,c'} \cdot \mathbf{1}\{y_{a''} = 1\} = p_s H_{sa''}$ $\forall s, a''$, where $\Phi \equiv \{\phi_{y_0,y_1,y_2,s,c} : y_0 \in \{0,1\}, y_1 \in \{0,1\}, y_2 \in \{0,1\}, s \in \mathcal{A}, c \in \mathcal{A}\}$.

The joint posterior of these parameters can be factorized as $f(\boldsymbol{H}, \boldsymbol{G}, \boldsymbol{p}, \boldsymbol{q}|\boldsymbol{X}) = f(\boldsymbol{H}|\hat{\boldsymbol{F}}^1)\, f(\boldsymbol{G}|\hat{\boldsymbol{F}}^0)$ $f(\boldsymbol{q}|\boldsymbol{n}^0_s) f(\boldsymbol{p}|\boldsymbol{n})$ under Assumptions 1 and 2. We use the improper priors $H_{sa} \sim \text{Beta}(0, 0)$ and $G_{sa} \sim \text{Beta}(0, 0)$. The posteriors are then given by $H_{sa} \sim \text{Beta}(\sum_{i=1}^N \mathbf{1}\{Y_i = 1, S_i = s, A_i = a, D_i = 1\}, \sum_{i=1}^N \mathbf{1}\{Y_i = 0, S_i = s, A_i = a, D_i = 1\})$ and $G_{sa} \sim \text{Beta}(\sum_{i=1}^N \mathbf{1}\{Y_i = 1, S_i = s, A_i = a, D_i = 0\}, \sum_{i=1}^N \mathbf{1}\{Y_i = 0, S_i = s, A_i = a, D_i = 0\})$.

A.6 Statistical Inference for the Sensitivity Analysis

Our approach to statistical inference for the sensitivity analysis in Section 5 is similar to the procedure outlined in Section 6. In addition to the parameters defined there, we have the naïve estimates $\boldsymbol{\eta} = \{\eta(a|s) : a, s \in \mathcal{A}\}$, where $\eta(a|s) = \mathbb{E}[Y_i|S_i = s, A_i = a, D_i = 1]$. For a given value of the sensitivity parameter, $\rho = \rho_{ac} = \rho_{a'c}$, the sets of upper and lower bounds on $\tau(a, a'|c)$ are denoted $\boldsymbol{\tau}^-_\rho$ and $\boldsymbol{\tau}^+_\rho$ for $a, a', c \in \mathcal{A}$.

Given $\boldsymbol{\pi}^-$, $\boldsymbol{\pi}^+$, $\boldsymbol{p}$, $\boldsymbol{q}$, and $\boldsymbol{\eta}$, we can deterministically find $\boldsymbol{\tau}^-_\rho$ and $\boldsymbol{\tau}^+_\rho$. Each pair of $\tau^-_\rho(a, a'|c)$ and

$\tau_\rho^+(a, a'|c)$ are equal to the endpoints of the following interval:

$$\begin{aligned}
&\tau(a, a'|c) \in \\
&\left(\left[\eta(a|c) - \rho_{ac},\ \eta(a|c) + \rho_{ac} \right] \bigcap \left[\sum_{s \in \mathcal{A}} \pi^-(a|s,c) \frac{q_{sc} p_s}{\sum_{s' \in \mathcal{A}} q_{s'c} p_{s'}},\ \sum_{s \in \mathcal{A}} \pi^+(a|s,c) \frac{q_{sc} p_s}{\sum_{s' \in \mathcal{A}} q_{s'c} p_{s'}} \right] \right) \\
&\quad - \left(\left[\eta(a|c) - \rho_{a'c},\ \eta(a|c) + \rho_{ac} \right] \bigcap \left[\sum_{s \in \mathcal{A}} \pi^-(a'|s,c) \frac{q_{sc} p_s}{\sum_{s' \in \mathcal{A}} q_{s'c} p_{s'}},\ \sum_{s \in \mathcal{A}} \pi^+(a'|s,c) \frac{q_{sc} p_s}{\sum_{s' \in \mathcal{A}} q_{s'c} p_{s'}} \right] \right).
\end{aligned}$$

We therefore simulate from the posterior of $(\boldsymbol{\tau}^-, \boldsymbol{\tau}^+)$ by drawing samples of $\boldsymbol{\pi}^-$, $\boldsymbol{\pi}^+$, $\boldsymbol{\eta}$, $\boldsymbol{p}$ and $\boldsymbol{q}$,

$$f(\boldsymbol{\pi}^+, \boldsymbol{\pi}^-, \boldsymbol{\eta}, \boldsymbol{p}, \boldsymbol{q} | \boldsymbol{X}) \ = \ f(\boldsymbol{\pi}^+, \boldsymbol{\pi}^-, \boldsymbol{\eta} | \hat{\boldsymbol{F}}^1, \hat{\boldsymbol{F}}^0, \boldsymbol{q})\ f(\boldsymbol{q} | \boldsymbol{n}_s^0)\ f(\boldsymbol{p} | \boldsymbol{n})$$

under Assumptions 1 and 2. Note that this differs from Appendix A.5 only in that the distributions of $\boldsymbol{\pi}^+$ and $\boldsymbol{\pi}^-$ are considered jointly with $\boldsymbol{\eta}$. These have the additional independence relations

$$\begin{aligned}
\eta(a|s) \ &\perp\!\!\!\perp \ \pi^+(a|s', c), \pi^-(a|s', c), \eta(a|s') \mid \hat{\boldsymbol{F}}^1, \hat{\boldsymbol{F}}^0, \boldsymbol{q} \quad \text{and} \\
\eta(a|s) \ &\perp\!\!\!\perp \ \pi^+(a'|s, c), \pi^-(a'|s, c), \eta(a'|s) \mid \hat{\boldsymbol{F}}^1, \hat{\boldsymbol{F}}^0, \boldsymbol{q}
\end{aligned}$$

for $s \neq s'$ and $a \neq a'$.

We can therefore approximate the posterior of sensitivity bounds by Monte Carlo simulation of $\boldsymbol{p}$, $\boldsymbol{q}$, and the trivariate distributions $[\pi^-(a|s,c),\ \pi^+(a|s,c),\ \eta(a|c)]$ for $a \in \{a', a''\}$ and $s \in \mathcal{A}$. By the Bayesian central limit theorem, the latter is given by

$$\begin{bmatrix} \pi^-(a|s,c) \\ \pi^+(a|s,c) \\ \eta(a|c) \end{bmatrix} \mid \boldsymbol{q}, \boldsymbol{X} \ \approx \ \text{Normal} \left(\begin{bmatrix} \bar{\pi}^-(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) \\ \bar{\pi}^+(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) \\ \bar{\eta}(a|c,\boldsymbol{X}) \end{bmatrix}, \boldsymbol{\Sigma}(a|s,c) \right), \quad \text{where}$$

$$\boldsymbol{\Sigma}(a|s,c) = \begin{bmatrix} V^-(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) & C(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) & C_\eta^-(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) \\ C(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) & V^+(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) & C_\eta^+(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) \\ C_\eta^-(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) & C_\eta^+(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) & V_\eta(a|s,\boldsymbol{q}_s,\boldsymbol{X}) \end{bmatrix}$$

when N is large, and the additional parameters $\bar{\eta}$, C_η^-, C_η^+, and V_η are defined below.

Note that naïve estimate $\eta(a|s)$ is point-identified, and its posterior mean and variance are equivalent to the sample mean and the sampling variance for the mean for the corresponding forced-choice units. These are given by:

$$\begin{aligned}
\bar{\eta}(a|s) &= \underline{y} + \int_{\underline{y}}^{\infty} 1 - F(y|s,a,1)\,\mathrm{d}y, \\
V_\eta(a|s) &= \frac{2}{n^1_{sa}} \int_{\underline{y}}^{\infty} \int_{y}^{\infty} F(y|s,a,1)\,(1 - F(x|s,a,1))\;\mathrm{d}x\mathrm{d}y.
\end{aligned}$$

Derivations closely follow Section A.5 and therefore are omitted here. Estimation can be done by plug-in with an optional small sample correction.

The posterior of $\eta(a|s)$ covaries with those of $\pi^-(a|s,c)$ and $\pi^+(a|s,c)$ because the latter parameters depend partially on the ECDF of the same forced-choice units.

$$\begin{aligned}
&C^-_\eta\left(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}\right) \\
&\quad= \mathrm{Cov}\left(\hat{\Gamma}^{-1}_a(b) - \int_{\underline{y}}^{\hat{\Gamma}^{-1}_a(b)} \frac{\hat{F}(y|s,a,1) - \hat{F}(y|s,a,0)q_{sa}}{q_{sc}}\,\mathrm{d}y, \underline{y} + \int_{\underline{y}}^{\infty} 1 - \hat{F}(y|s,a,1)\,\mathrm{d}y\right) \\
&\quad= \frac{2}{n^1_{sa}\cdot q_{sc}} \int_{\underline{y}}^{\hat{\Gamma}^{-1}_a(b)} \int_{y}^{\hat{\Gamma}^{-1}_a(b)} F(y|s,a,1)\,(1 - F(x|s,a,1))\;\mathrm{d}x\mathrm{d}y \\
&\qquad+ \frac{1}{n^1_{sa}\cdot q_{sc}} \int_{\underline{y}}^{\hat{\Gamma}^{-1}_a(b)} \int_{\hat{\Gamma}^{-1}_a(b)}^{\infty} F(y|s,a,1)\,(1 - F(x|s,a,1))\;\mathrm{d}x\mathrm{d}y \\
&C^+_\eta\left(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}\right) \\
&\quad= \mathrm{Cov}\left(\hat{\Gamma}^{-1}_a(1-b) - \int_{\hat{\Gamma}^{-1}_a(1-b)}^{\infty} \frac{q_{sa} + \hat{F}(y|s,a,1) - \hat{F}(y|s,a,0)q_{sa} - 1}{q_{sc}}\,\mathrm{d}y,\right. \\
&\qquad\qquad\left. \underline{y} + \int_{\underline{y}}^{\infty} 1 - \hat{F}(y|s,a,1)\,\mathrm{d}y\right) \\
&\quad= \frac{1}{n^1_{sa}\cdot q_{sc}} \int_{\underline{y}}^{\hat{\Gamma}^{-1}_a(1-b)} \int_{\hat{\Gamma}^{-1}_a(1-b)}^{\infty} F(y|s,a,1)\,(1 - F(x|s,a,1))\;\mathrm{d}x\mathrm{d}y \\
&\qquad+ \frac{2}{n^1_{sa}\cdot q_{sc}} \int_{\hat{\Gamma}^{-1}_a(1-b)}^{\infty} \int_{y}^{\infty} F(y|s,a,1)\,(1 - F(x|s,a,1))\;\mathrm{d}x\mathrm{d}y
\end{aligned}$$

for any $s, c \neq a \in \mathcal{A}$.

Thus, each draw of the sensitivity results from their posterior is generated by the following procedure:

1. Draw $\boldsymbol{p} \equiv [p_s] \sim \text{Dirichlet}(\boldsymbol{n})$, where $\boldsymbol{n} \equiv [n_s] = \left[\sum_{i=1}^N \mathbf{1}\{S_i = 0\}, \cdots, \sum_{i=1}^N \mathbf{1}\{S_i = J-1\}\right]^\top$.

2. For each $s \in \mathcal{A}$:

   (a) Draw $\boldsymbol{q}_s \equiv [q_{sa}] \sim$ Dirichlet($\boldsymbol{n}_s^0$), where $\boldsymbol{n}_s^0 \equiv [n_{sa}^0] = \left[\sum_{i=1}^N \mathbf{1}\{S_i = s, A_i = 0, D_i = 0\}, \cdots, \sum_{i=1}^N \mathbf{1}\{S_i = s, A_i = J-1, D_i = 0\}\right]^\top$;

   (b) For each a and $c \in \mathcal{A}$, draw a triplet $[\pi^-(a|s,c), \pi^+(a|s,c), \eta(a|s)]$ from the trivariate normal distribution defined above.

3. For a given ρ, calculate a simulated draw of $[\tau_\rho^-(a,a'|c), \tau_\rho^+(a,a'|c)]$ according to equation (9).

The sensitivity procedure for binary outcomes differs only in the last two steps:

2. (b) For each $a \in \mathcal{A}$, draw H_{sa} and G_{sa} from the posteriors discussed in Sections 6 and A.5.

3. Calculate a simulated draw of $[\tau^-(a,a'|c), \tau^+(a,a'|c)]$ by solving the linear programming problem in equation (24), with the additional sensitivity constraints $\sum_{y_0\in\{0,1\}} \sum_{y_1\in\{0,1\}} \sum_{y_2\in\{0,1\}} \sum_{s\in\mathcal{A}} \phi_{y_0,y_1,y_2,s,c} \mathbf{1}\{y_{a^*} = 1\} \geq (H_{sa^*} - \rho_{a^*c}) \sum_{s\in\mathcal{A}} q_{sc} p_s$ and $\sum_{y_0\in\{0,1\}} \sum_{y_1\in\{0,1\}} \sum_{y_2\in\{0,1\}} \sum_{s\in\mathcal{A}} \phi_{y_0,y_1,y_2,s,c} \mathbf{1}\{y_{a^*} = 1\} \leq (H_{sa^*} + \rho_{a^*c}) \sum_{s\in\mathcal{A}} q_{sc} p_s$ for given c and $a^* \in \{a, a'\}$.

A.7 Additional Simulation Results

In this section, we present additional results from the simulations described in Section 8. First, we explore the performance of the EM-algorithm-based parametric approach proposed by Long et al. (2008) (hereafter LLL) in a setting close to our empirical application. This necessitates extending LLL's original methodology, as it was developed for a binary treatment. We thus modify their parametric model to accommodate a categorical treatment by modeling the treatment choice with the multinomial logit model, as opposed to the binary logit model. (We have confirmed that our own R implementation of this extension replicates the simulation results reported by LLL in their original article almost exactly.) To make LLL's approach comparable to our proposed method in terms of observed information used, we set subjects' stated preferences as the covariate in their choice and outcome models (i.e., $X_{1i} = X_{2i} = S_i$ using their notation). We then apply the LLL estimator to the same 500 simulated datasets as in Section 8.

	CD=0.00	CD=0.33	CD=0.67	CD=1.00
LLL	0.053	0.028	0.019	-0.020
naïve	0.002	0.011	0.023	0.038
min	-0.001	-0.001	-0.001	0.000
max	-0.001	-0.001	-0.001	-0.001

Table A.1: LLL bias for various CD values, holding OD at zero. Naïve and bounds biases from Section 8.1 are reproduced here for convenience.

	OD=0.00	OD=0.33	OD=0.67	OD=1.00
LLL	0.053	0.062	0.072	0.080
naïve	0.002	0.011	0.020	0.030
min	-0.001	0.001	0.001	0.001
max	-0.001	-0.002	-0.002	-0.001

Table A.2: LLL bias for various OD values, holding CD at zero. Naïve and bounds biases from Section 8.2 are reproduced here for convenience.

Tables A.1 and A.2 show the results in terms of bias at the sample size of 3,000 (second row from the top), along with the comparable results for the naïve estimator and our proposed bounds estimator (third row and below), which are reproduced from Tables 2 and 3 in the main text. Somewhat surprisingly, and contrary to the original findings by LLL based on a much simpler simulation setup, the LLL estimator exhibits substantial bias even when both CD and OD are zero. This suggests that finite-sample performance of the LLL estimator is rather poor when applied to datasets like ours, rendering it an unattractive option for inference.

Next, we contrast the proposed Bayesian inferential approach described in Section A.5 to an alternative method based on the nonparametric bootstrap. We construct the 95% bootstrap confidence intervals by taking the 2.5th and 97.5th percentiles of parameter estimates in 1000 bootstrap draws. For the bounds, we take those percentiles from the lower and upper bound estimates, respectively, to construct confidence intervals that are purported to cover the nonparametric bounds 95% of the time.

Tables A.3 and A.4 show estimated coverage rates for the 95% bootstrap confidence intervals at various values of the CD and OD parameters. The comparable results for our proposed Bayesian intervals can be found in Tables A.3 and A.4 in the main text. In general, we find that the coverage of the bootstrap intervals is noticeably below that of our proposed method, and the bootstrap coverage rates are

n	CD=0.00	CD=0.33	CD=0.67	CD=1.00
500	0.944	0.930	0.895	0.891
1000	0.941	0.915	0.911	0.906
3000	0.952	0.914	0.908	0.924
10000	0.930	0.924	0.924	0.928
50000	0.936	0.940	0.940	0.942

Table A.3: Bootstrap coverage rates for various CD values, holding OD at zero.

n	OD=0.00	OD=0.33	OD=0.67	OD=1.00
500	0.944	0.954	0.949	0.950
1000	0.941	0.960	0.956	0.959
3000	0.952	0.942	0.952	0.944
10000	0.930	0.952	0.950	0.958
50000	0.936	0.948	0.944	0.946

Table A.4: Bootstrap coverage rates for various OD values, holding CD at zero.

substantially below nominal at lower sample sizes and for larger values of the CD parameter.

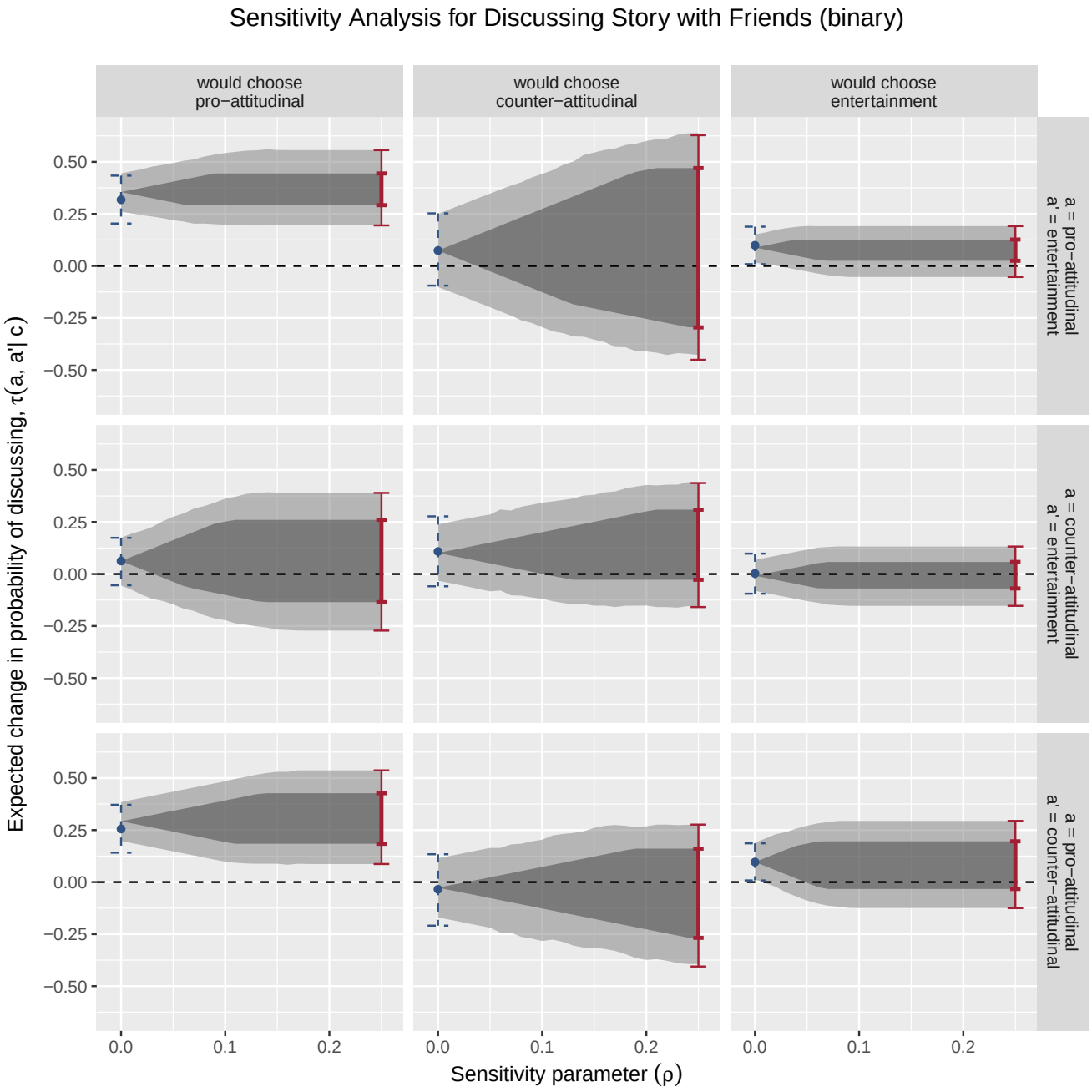

Figure A.1: Sensitivity Analysis for the ACTE of Partisan News Media (Binary Outcome). The plots correspond to the right panel of Figure 2. See caption for Figure 3 for the explanation of graph elements.