Design, Identification, and Sensitivity Analysis for Patient Preference Trials[*]

Dean Knox[†] Teppei Yamamoto[‡] Matthew A. Baum[§] Adam Berinsky[¶]

First Draft: July 20, 2014
This Draft: September 1, 2016

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 overall 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. Finally, we 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.

Key Words: randomized controlled trials, external validity, causal inference, nonparametric bounds, stated and revealed preferences

[*] 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.

[†] Ph.D. Candidate, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: dcknox@mit.edu, URL: www.dcknox.com

[‡] Associate Professor, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: teppei@mit.edu, URL: http://web.mit.edu/teppei/www

[§] Professor, John F. Kennedy School of Government, Harvard University.

[¶] Professor, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139.

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 the 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 render the study externally invalid, if researchers are not only interested in simple average treatment effects but also in broader implications of their empirical findings. For example, a psychiatric 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, such nuanced inference cannot be made because all subjects are forced to take treatments randomly chosen by the researcher.

In this paper, 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 that have previously been 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 are asked to state their preferred treatments prior to entering the study. Then, they are randomized into either a standard RCT or a self-selection condition, where they are allowed to choose the treatment as they would in the real world. Finally, the outcome variables of interest are measured. The proposed design is novel in that it allows the researcher to estimate how accurately stated preferences predict the actual choice of treatments. In the social sciences, it is a widely shared concern that respondents to a survey question may not accurately report their underlying preferences to the interviewer (whether consciously or subconsciously) and their tendency to do so may be systematically correlated with unobserved characteristics that interact with the treatment effects.

Using the potential outcomes framework of causal inference (Neyman, 1923; Rubin, 1974), we formally define a causal quantity which represents the conditional average treatment effect for a subpopulation of subjects who would choose a particular treatment option. We show that the point identification of this quantity for a multi-valued treatment requires the strong assumption that the discrepancy between stated preference and actual choice is ignorable. Then, without making this assumption, we derive nonparametric sharp bounds on this causal quantity. Finally, we propose a sensitivity analysis where we quantify the assumed informativeness of the stated preferences about revealed preferences via a sensitivity parameter and analyze how the quantity of interest responds to the change in this parameter. To illustrate the proposed design and methodology, we implement them in an original survey experiment where we investigate the effect of partisan political news media on the subjects' perception about media and political behavior. Our primary interest is in how the effect varies depending on whether they would actually consume such partisan media if they could choose.

Despite the prevalence of preference trials across scientific disciplines, very few methodological investigations have been conducted on the topic from the percpective of causal inference. A notable exception is Long, Little and Lin (2008), who employ a framework similar to ours to define the 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 any distributional or functional-form assumptions for better credibility of the resulting inference.

The rest of the paper proceeds as follows. Section 2 describes the background motivation of the empirical example. Section 3 introduces the notation and defines causal quantities of interest and assumptions. Sections 4 and 6 discuss the proposed methodology. Section 7 applies the method to the empirical example. Section 8 concludes.

A Motivating Example

In this section, we provide background information on an original randomized experiment where we implemented the proposed PPT design to examine the effects of partisan news media on political choice.

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 research has sought to delineate the effects of consuming ideologically polarized media on attitudes towards the media in general. 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 percent. From a normative perspective, the worry is that people who distrust the media will conclude it cannot report in an unbiased manner and so dismiss as unreliable its content. As a result, the public may increasingly become suspicious of and antagonistic toward the news media more generally (Arceneaux, Johnson and Murphy, 2012; Ladd, 2012). Such attitudes, in turn, may have implications for political behavior.

To explore this phenomenon, we conducted an experiment in June 2014 on a sample of 3,023 American adults, recruited by Survey Sampling International (SSI). Our goal was to estimate the effect of exposing these subjects to pro- and counter-attitudinal political news programming (as opposed to non-political entertainment shows) on their sentiment towards specific news programs and the media in general. We also explored whether such programming produces behavioral responses, such as changes in propensity to discuss it 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). The two political shows — Rachel Maddow and The O'Reilly Factor — are then coded as either pro- or counter-attitudinal for each subject based on their party identification (Democratic or Republican). These two clips are carefully selected 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, the two entertainment shows were merged 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 may 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. Results from this experiment will be analyzed with our proposed methodology and presented in Section 7.

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.

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 $A_i \in \mathcal{A} \equiv \{0, 1, ..., J-1\}$ 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 patient preference trials 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 below, 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. 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 $D_i = 1$. For the subjects with $D_i = 0$, their treatments equal the treatments they have chosen, which we denote by $C_i \in \mathcal{A}$. Therefore, we have

$A_i = C_i$ if $D_i = 0$. Finally, the outcome of interest 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, i.e., the potential outcomes remain stable across the two design conditions. Long, Little and Lin (2008) call 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.We use Y_i to denote the observed outcome of subject i. By definition, 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 cumulative distribution function (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, such as social desirability bias. Thus, a "naïve" analysis that takes the stated preferences at their face value and ignores the possible measurement error may lead to an estimate that shows unrealistically high degree of certainty, as we illustrate with the media choice example in Section 7. Finally, note that we allow the treatment variable to be multi-valued, instead of binary. In fact, as previously shown by Long, Little

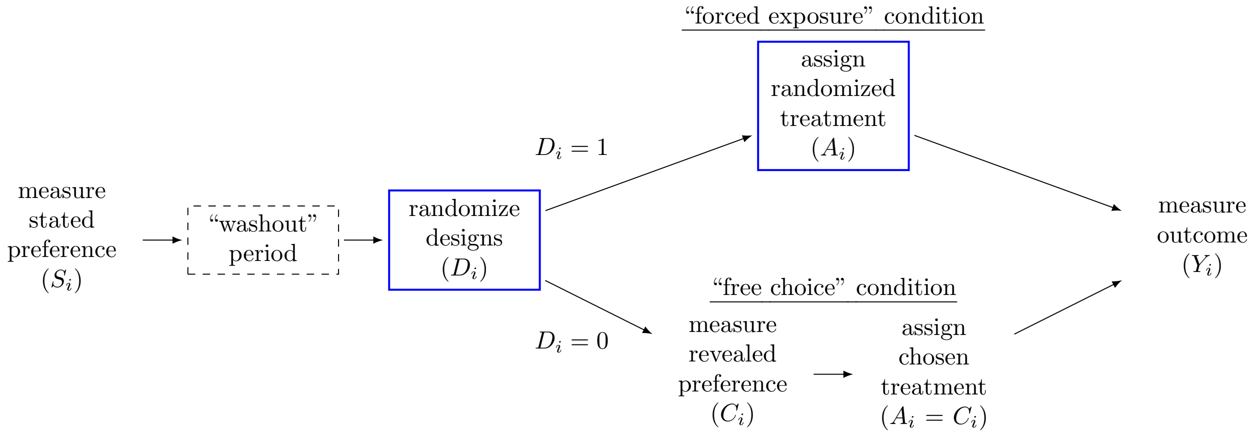
"forced exposure" condition
assign randomized treatment (A_i)
$D_i = 1$
measure stated preference (S_i)
"washout" period
randomize designs (D_i)
measure outcome (Y_i)
"free choice" condition
$D_i = 0$
measure revealed preference (C_i)
assign chosen treatment ($A_i = C_i$)

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.

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 average choice-specific treatment effects (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 numerous previous studies in both social and medical sciences that utilize designs closely related to ours. However, as far as we are aware, no other study combines the measurement of stated preferences with randomization into either the forced exposure or free choice condition (King et al., 2005), which we regard as important. For 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. However, because these two studies are conducted separately on populations with possibly different characteristics, it is not straightforward to make inference from combined data.

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 non-zero 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 political 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 choice.

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

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

for any a, a' and $c \in \mathcal{A}$. We call this quantity the *average choice-specific treatment effect (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 *average treatment effect 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 paper, we opt to introduce the new causal quantity ACTE because its interpretation is clearer and less affected by auxiliary design assumptions than the ATT.

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)

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

Long, Little and Lin (2008) refer 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)

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

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)

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

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 will 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, 2 and 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, 2 and 3.

$$\underline{y} \leq \frac{\mathbb{E}[Y_i|C_i = a, D_i = 0] - \mathbb{E}[Y_i|C_i = S_i = a, D_i = 0] \Pr(C_i = a|S_i = a, D_i = 0)}{1 - \Pr(C_i = a|S_i = a, D_i = 0)} \leq \overline{y} \tag{3}$$

for any $a \in \mathcal{A}$. Proofs are provided in Appendix A.1.

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 reseacher 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.

Nonparametric Identification Analysis

In this section, we present the results of our nonparametric identification analysis for the ACTE 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 Appendix A.2, we show that the ACTE can be expressed as follows under those assumptions.

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

for any a, a' and $c \in \mathcal{A}$. Equation (4) immediately gives us three important results. First, equation (4) contains a total of at least $J - 2$ terms (when $a \neq a' = c$ or $a = c \neq a'$) and 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. 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 follows.

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

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 that they only consider the case of $J = 2$. The same result is also obtained by Long, Little and Lin (2008) using a framework more similar to ours. Thus, we verify these 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, 2 and 3 and given by the following expression.

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

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.182–3) 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 context, 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 are not willing to make Assumption 3 or restrict the analysis to binary treatments? Here, we present two partial identification results, which provide *sharp bounds* (i.e., the tightest possible given all the observed information; Manski, 1995) on $\tau(a, a'|c)$ under alternative scenarios.

General Results for Unbounded Outcomes

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

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

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:

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

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

These bounds are obtained 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, the sharp bounds on $\mathbb{E}[Y_i(a) \mid C_i = c]$ can be derived from these bounds, which can then be used to construct bounds on

$\tau(a, a'|c)$. A detailed proof can be found in Appendix A.3.

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 $a' = c$, i.e., 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.

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 achived via the linear programming approach based on principal stratification (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 $(Y_i(0), ..., Y_i(J-1))$ 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 $\phi_{y_0,...,y_{J-1},s,c} \equiv \Pr(Y_i(0) = y_0, ..., Y_i(J-1) = y_{J-1}, S_i = s, C_i = c)$, where $y_0, ..., y_{J-1} \in \{0, 1\}$ 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 non-binary categorical outcomes with a straightforward extension, which we do not pursue in the current paper in order 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.*

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

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

A proof is provided in Appendix A.4. 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. For the $\tau(a, c|c)$ case with binary outcomes, where equation 7 simplifies to

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

we numerically verify that both sets of sharp bounds coincide, as they should.

Uncertainty Estimation

In this section, we discuss our methods for performing statistical inference for the large-sample bounds presented in Section 4. An important challenge in performing inference for partially identified parameters is that standard resampling methods such as the nonparametric bootstrap cannot be directly applied unless certain smoothness conditions are met (Horowitz, 2001). Here, we instead use 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. Specifically, we use the following procedure to obtain one simulated draw of the bounds, $\tau^-(a, a'|c)^*$ and $\tau^+(a, a'|c)^*$, from their posterior:

1. Draw $\boldsymbol{p} \equiv [p_s] \sim$ 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 $\boldsymbol{n}_s^0 \equiv [n_{sa}^0] = \Big[\sum_{i=1}^N \mathbf{1}\{S_i = s, A_i = 0, D_i = 0\}, \cdots, \sum_{i=1}^N \mathbf{1}\{S_i = s, A_i = J-1, D_i = 0\}\Big]^\top$;
   (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 Appendix A.5.
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 2b will be a draw from the univariate normal distribution with mean $\bar{\pi}$ and variance V, which are also given in Appendix A.5. A notable feature of this procedure is that it can approximate the posterior for the bounds on $\tau(a, a'|c)$ under a minimal assumption about the true distribution of the potential outcomes when sample size is large. The only parametric assumption for the procedure is the use of noninformative Dirichlet priors for $\boldsymbol{p}$ and $\boldsymbol{q}$. Details are provided in Appendix A.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 which is defined as,

$$\rho \;\equiv\; \max\left\{\left|\mathbb{E}[Y_i(a)|S_i = c] - \mathbb{E}[Y_i(a)|C_i = c]\right| : a, c \in \mathcal{A}, a \neq c\right\}.$$

In words, ρ represents the maximum absolute difference we allow to exist between the average potential outcome among those who state a particular treatment preference and the unobserved average potential outcome among those who actually choose that treatment. This definition of ρ implies the following additional constraint,

$$\mathbb{E}[Y_i|S_i = c, A_i = a, D_i = 1] - \rho \;\leq\; \mathbb{E}[Y_i(a)|C_i = c] \;\leq\; \mathbb{E}[Y_i|S_i = c, A_i = a, D_i = 1] + \rho, \qquad (8)$$

for all a and $c \in \mathcal{A}$ such that $a \neq c$.

The proposed sensitivity analysis proceeds by incorporating equation (8) to the calculation of bounds. For the general case of unbounded outcomes, an analytical solution becomes intractable because equation (8) can constrain unobserved conditional averages of potential outcomes via many interrelated inequality restrictions. We therefore find bounds on $\tau(a, a'|c)$ for a given ρ numerically by solving the following linear programming problem.

$$\min_{\Pi} \quad \text{and} \quad \max_{\Pi} \quad \sum_{s \in \mathcal{A}} \{\pi(a|s,c) - \pi(a'|s,c)\} \Pr(S_i = s | A_i = c, D_i = 0),$$

s.t.

$$\begin{aligned} &\lim_{y^* \to -\infty} \left[\int_{y^*}^{\infty} 1 - \min\left\{1, \frac{F(y|s,a^*,1) - F(y|s,a^*,0) P(a^*|s,0)}{\Pr(A_i = c|S_i = s, D_i = 0)}\right\} \mathrm{d}y + y^* \right] \leq \pi(a^*|s,c) \\ &\leq \lim_{y^* \to -\infty} \left[\int_{y^*}^{\infty} 1 - \max\left\{0, 1 - \frac{1 - F(y|s,a^*,1) + \{1 - F(y|s,a^*,0)\} P(a^*|s,0)}{\Pr(A_i = c|S_i = s, D_i = 0)}\right\} \mathrm{d}y + y^* \right] \end{aligned} \tag{9}$$

and

$$\mathbb{E}[Y_i|S_i = c, A_i = a^*, D_i = 1] - \rho \leq \sum_{s \in \mathcal{A}} \pi(a^*|s,c) \Pr(S_i = s|A_i = c, D_i = 0) \leq \mathbb{E}[Y_i|S_i = c, A_i = a^*, D_i = 1] + \rho \tag{10}$$

for $a^* \in \{a, a'\}$ and $s \in \mathcal{A}$, where $\Pi \equiv \{\pi(a^*|s,c) : a^* \in \{a,a'\}, s \in \mathcal{A}\}$. The solution to this problem corresponds to the upper and lower bounds on $\tau(a,a'|c)$, since the objective function equals $\tau(a,a'|c)$ when $\pi(a^*|s,c) = \mathbb{E}[Y_i(a^*)|S_i = s, C_i = c]$. Equations (9) and (10) are restrictions implied by Assumptions 1 and 2 (see the proof of Proposition 1 in Appendix A.3) and by equation (8), respectively. Again, when $a' = c$, the objective function can be simplified as $\sum_{s \in \mathcal{A}} \pi(a|s,c) \Pr(S_i = s|A_i = c, D_i = 0) - \mathbb{E}[Y_i|A_i = c, D_i = 0]$ and the resulting bounds are now sharp for a given value of ρ. Uncertainty estimates can be obtained for a given ρ via a Monte Carlo procedure analogous to the case of the bounds. Details are provided in Appendix A.6.

For binary outcomes, the sharp bounds can be numerically obtained for any $\tau(a,a'|c)$ and a given value of ρ 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 $\sum_{y_0 \in \{0,1\}} \sum_{y_1 \in \{0,1\}} \sum_{y_2 \in \{0,1\}} \sum_{s' \in \mathcal{A}} \phi_{y_0,y_1,y_2,s',c'} \mathbf{1}\{y_{a''} = 1\} \geq (\Pr(Y_i = 1 \mid S_i = c', A_i = a'', D_i = 1) - \rho) \Pr(C_i = c')$ and $\sum_{y_0 \in \{0,1\}} \sum_{y_1 \in \{0,1\}} \sum_{y_2 \in \{0,1\}} \sum_{s' \in \mathcal{A}} \phi_{y_0,y_1,y_2,s',c'} \mathbf{1}\{y_{a''} = 1\} \leq (\Pr(Y_i = 1 \mid S_i = c', A_i = a'', D_i = 1) + \rho) \Pr(C_i = c')$ for all $c', a'' \in \mathcal{A}$.

Empirical Application

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

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, we asked all subjects their preferences over the four television programs (listed in Section 2) early in the survey. Specifically, we asked, "If you were given the choice of the following four television programs to watch, which would you choose?" and we presented each choice with an accompanying screenshot of the host of the show, with the order of the shows being randomized.

Subsequently, we included a "washout" period in which subjects are asked various questions not directly related to the media choice (e.g. demographics, unrelated psychological experiments), including a question about their partisanship that we used to categorize their media preferences as pro- or counter-attitudinal. A primary purpose of inserting these filler questions for our study was to minimize the possibility that the measurement of stated preference might contaminate their voluntary choice of a television program in the free choice condition. Incorporating this kind of distractor questions (or even recontacting the subjects at a later time if feasible) might be an important practical element of the proposed PPT design to further enhance its external validity. After excluding subjects who were neither Democrat or Republican, 31% of the sample expressed a preference for pro-attitudinal media ($S_i = 1$), 12% for counter-attitudinal media ($S_i = -1$), and the remaining 57% for an entertainment show ($S_i = 0$).

Next, subjects were randomized with equal probability into the forced exposure ($D_i = 1$) and free choice ($D_i = 0$) conditions. Those in the forced choice arm were randomly assigned to watch pro-attitudinal media ($A_i = 1$), counter-attitudinal media ($A_i = -1$), or a randomly chosen entertainment program ($A_i = 0$), each with probability $1/3$. Those in the free choice arm were instead asked the question, “Which of these programs would you like to watch now?" with the same four options presented as before. Based on their partisanship and response, the actual choice C_i was recorded as 1, -1, or 0. 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 ($\Pr(C_i = 0|S_i = 0) = 0.91$, whereas $\Pr(C_i = 1|S_i = 1) = 0.81$ and $\Pr(C_i = -1|S_i = -1) = 0.77$). These subjects were assigned to view their choice, so that $A_i = C_i$ in the free-choice arm.

We consider two outcome variables. First, after viewing the program, respondents were asked to rate the clip they watched on a number of dimensions, which were summarized into an index of sentiment toward media. The index ranged between 0 and 1 and the mean and standard deviation were 0.61 and 0.17, respectively. Second, to gauge behavioral responses, subjects were asked how likely they would be to discuss the clip with a friend, which was summarized into a binary indicator. Overall, 62.5% of subjects were at least somewhat likely to discuss the viewed program.

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 media. For example, among those respondents in the free-choice group who stated a preference for pro-attitudinal media and also chose that program, mean sentiment was 0.67. In contrast, responses were significantly lower (by .07) among free-choice units that stated a preference for entertainment but actually chose pro-attitudinal media.

Nonparametric Bounds

Given the evidence that stated preferences of subjects do not accurately reflect their actual choice, 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% confidence intervals obtained via the nonparametric bootstrap (Horowitz and Manski, 2000). The left panel presents results for subjects’ sentiment toward the media watched (continuous; 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 de-

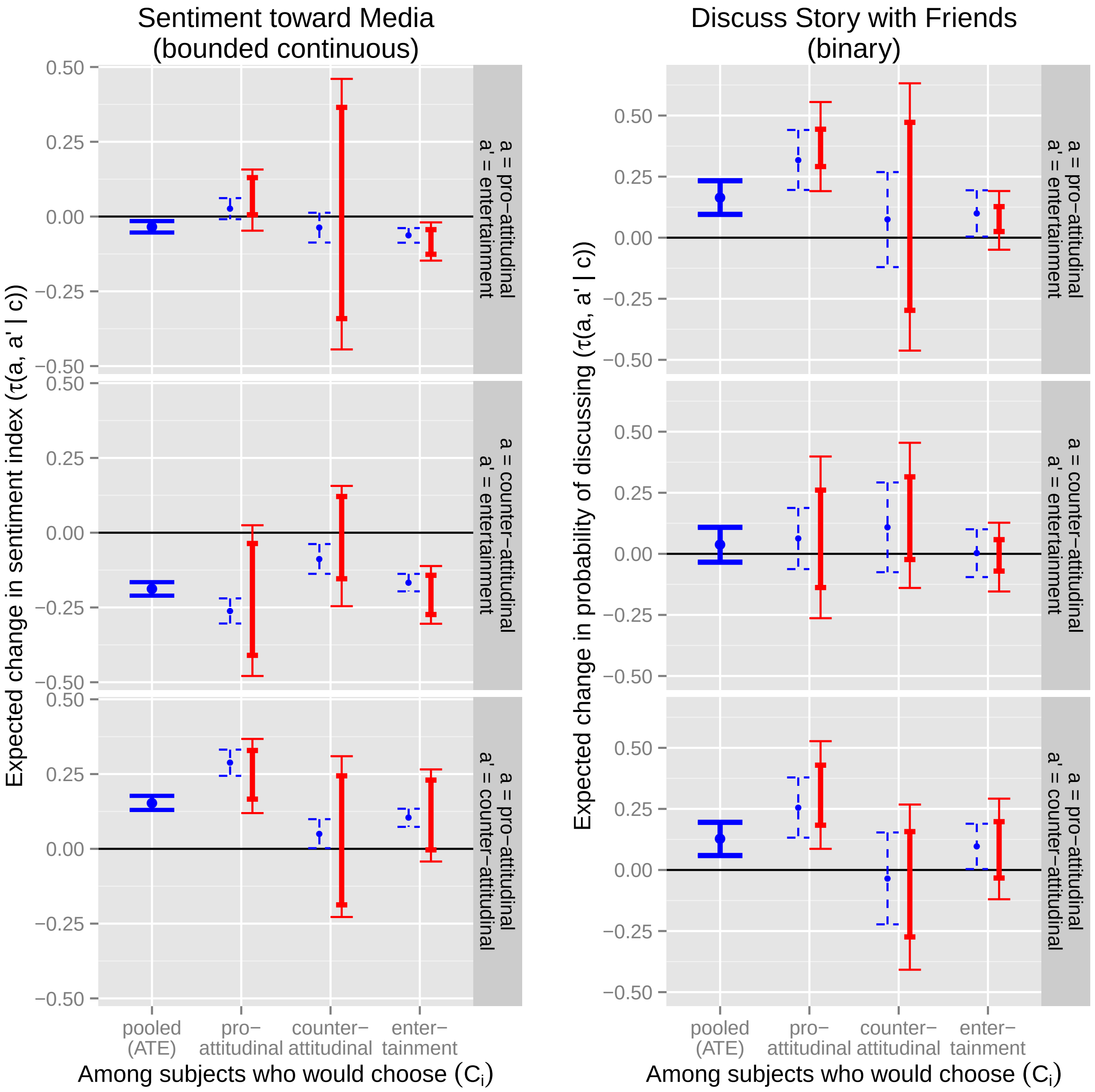

Figure 2: Estimated Nonparametric Bounds on the ACTE of Partisan News Media. Vertically stacked plots correspond to the same outcome variable. Horizontally aligned plots depict the effect of a particular change in the assigned media, i.e., $\mathbb{E}[Y_i(a) - Y_i(a')|C_i = c]$. Pairs of lines correspond 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 dotted error bars representing 95% asymptotic confidence intervals. Solid thick red error bars are estimated bounds and thin error bars give 95% bootstrap confidence intervals.

Free-choice Condition ($D_i = 0$)

Stated Preference (S_i)		1			-1			0		
Actual Choice ($C_i = A_i$)		1	-1	0	1	-1	0	1	-1	0
Strata proportions		.25	.02	.03	.01	.09	.02	.03	.02	.53
Outcomes (Y_i)	Sentiment toward media	.67	.51	.66	.52	.57	.60	.60	.54	.68
	Likely to discuss	.78	.76	.63	.62	.77	.68	.85	.80	.59

Forced-exposure Condition ($D_i = 1$)

Stated Preference (S_i)		1			-1			0		
Randomized Treatment (A_i)		1	-1	0	1	-1	0	1	-1	0
Strata proportions		.10	.11	.11	.04	.05	.05	.20	.18	.17
Outcomes (Y_i)	Sentiment toward media	.67	.38	.64	.59	.54	.63	.57	.47	.64
	Likely to discuss	.76	.50	.44	.73	.78	.67	.68	.57	.58

Table 1: Summary of Observed Data in the Media Choice Experiment. 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.

picts the effect of a particular change in the assigned media, from pro-attitudinal to entertainment (top), counter-attitudinal to entertainment (middle) and pro-attitudinal to counter-attitudinal (bottom). The left-most blue solid circle (point estimate) and arrow (95% asymptotic confidence itnerval) 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, 2 and 3, i.e., the naïve estimates that assume the ignorability of the discrepancy between stated preferences and actual choices. Blue dotted 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 and thin lines representing bootstrap confidence intervals.

For example, consider the middle bars in the center left plot. Here, blue dotted 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 no-assumption 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), but the difference between pro- and counter-attitudinal media among this group is not 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 (versus entertainment, center left) but these bounds are not 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 (versus entertainment, top right) and entirely inconclusive (versus counter-attitudinal media, bottom right).

Sensitivity Analysis

Next, we apply the sensitivity analysis developed in Section 6 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 analysis for the sentiment index. The results are presented in Figure 3.

Using bounds on mean strata potential outcomes (not presented), we find that the estimated maximal difference for any strata is 0.225; 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 0.12 can be ruled out conclusively. We thus view $\rho = 0.12$ as a fairly high value, equivalent to roughly three-quarters of a standard deviation in the outcome variable. Sensitivity results are not shown for $\rho < .05$, because in this region, it becomes impossible to simultaneously satisfy the constraints implied by ρ and the naïve results, on the one hand, and the bounding constraints, on the other. Thus, neglecting sampling error, the true value of ρ should lie somewhere in $[0.05, 0.225]$.

For illustration we focus on the center row of Figure 3, where 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, even before the minimum possible $\rho = 0.05$ is reached. Estimated bounds include zero for values of $\rho > 0.074$, less than half of a standard deviation in the outcome variable. In contrast, the naïve result that counter-attitudinal media negatively affects media sentiment among those that would in fact prefer to watch pro-attitudinal media (center left) provides an example of a relatively robust finding. The 95% bounds confidence interval does not span zero until the fairly high value of $\rho = 0.115$, and the estimated bounds themselves remain negative even when no assumptions are made about the informativeness of stated preferences.

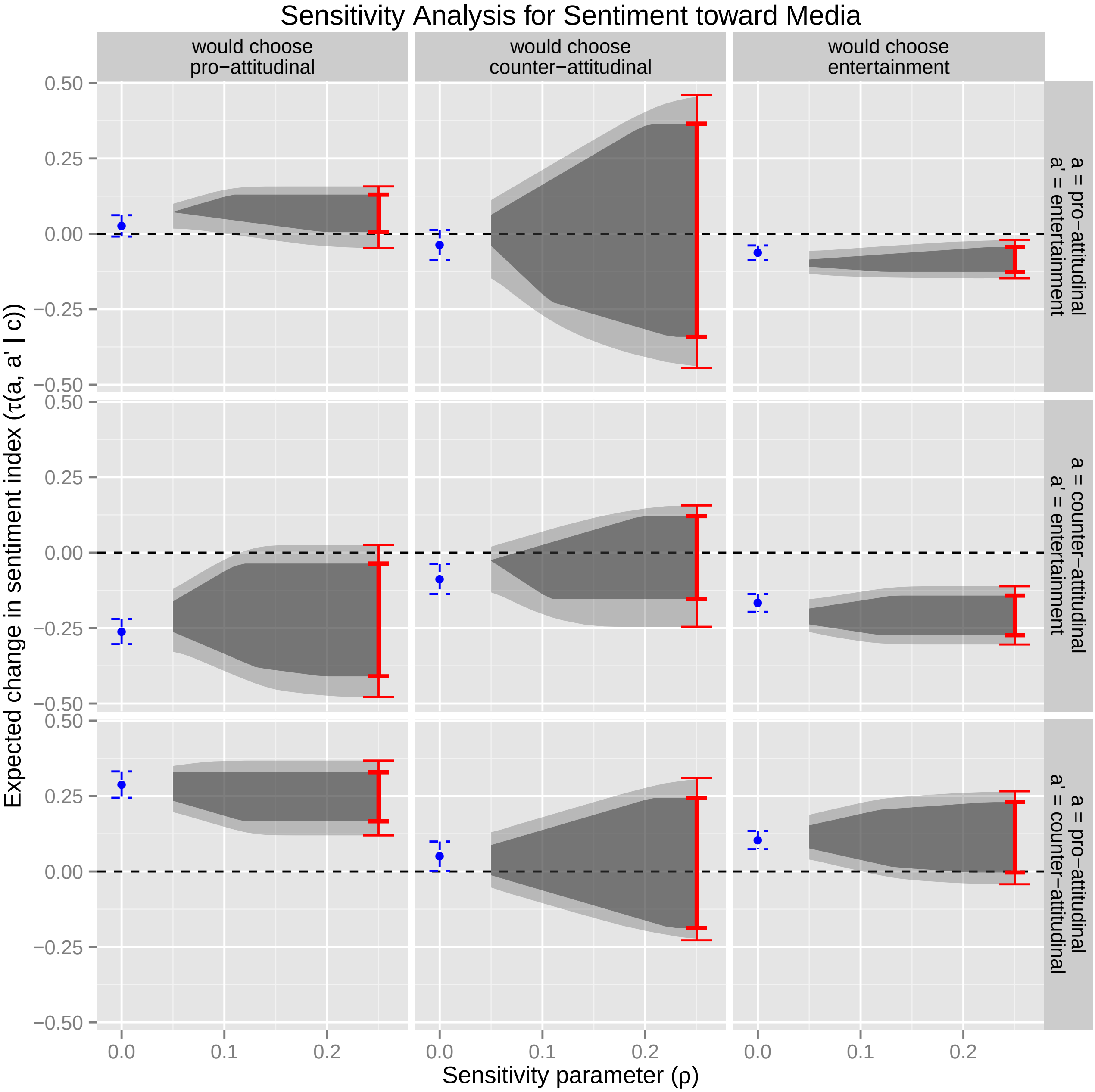

Figure 3: Sensitivity Analysis for the ACTE of Partisan News Media. The plots correspond to the left panel of Figure 2. On the left side of each plot, a blue point and error bars represent the naïve estimate and 95% asymptotic confidence intervals, respectively. On the right side, thick red error bars represent no-assumption bounds and thin red error bars represent 95% bootstrap confidence intervals. The dark shaded region between these depicts how bounds grow narrower as additional information from the naïve estimates are incorporated (ρ grows small). Lightly shaded regions are 95% bootstrap confidence regions for sensitivity results.

Concluding Remarks

Scholars of social and medical sciences have long sought to enhance the external validity of randomized experiments by various means. PPTs have often been adopted in medical research to incorporate the preferences of experimental subjects over treatment options into the study design, thereby tackling the question of whether treatments have impacts on the type of units 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, the potential discrepancy between subjects' stated and revealed preferences has been largely neglected in the existing literature.

In this paper, 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 difference between stated and revealed preferences via nonparametric identification analysis and sensitivity analysis. As we illustrate in an original empirical example where we implement 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 any untestable assumptions.

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 paper. A major motivation for PPTs in medical research is the concern that patients who strongly prefer one treatment option to others may not follow experimental protocols and cross over to another treatment arm or dropping out of the study, damaging the internal validity of the experiment. One natural direction for future research is, therefore, to incorporate such complications under the current framework.

Appendix

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,

$$\mathbb{E}[Y_i(a)|C_i = a] \;=\; \mathbb{E}[Y_i(a)|S_i = a], \tag{11}$$

for all $a \in \mathcal{A}$. This relationship directly implies equation (2) under Assumptions 1 and 2. Second, note that equation (11) also implies,

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

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

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

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

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

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

and $J-2$ terms remain unidentified.

Proof of Proposition 1

We begin by establishing several lemmas.

Lemma A.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,

$$\Gamma_a^+(y,c|s,a) = \min\left\{\Gamma_a(\infty,c|s,a), \Gamma_a(y,\infty|s,a)\right\}, \tag{12}$$

$$\Gamma_a^-(y,c|s,a) = \max\left\{0, \Gamma_a(\infty,c|s,a) + \Gamma_a(y,\infty|s,a) - 1\right\}. \tag{13}$$

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 (12) and (13) yields the results in Lemma A.1. ▮

Lemma A.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

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

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 $\Gamma_a^*(y,0\mid s,a)$,

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

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 A.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,*

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

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 A.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 A.1, these bounds simplify when $c=0$ to

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

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

and

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

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 A.2. Now, note that $\tau(a, a'|0)$ can be written under Assumption 1 as,

$$\tau(a,a'|0) = \sum_{s\in\mathcal{A}} \{\pi(a|s,0)-\pi(a'|s,0)\}\Pr(S_i=s|A_i=0,D_i=0), \tag{14}$$

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

$$\pi(0|s,0) = \mathbb{E}[Y_i|A_i=0,S_i=s,D_i=0], \tag{15}$$

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

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

$$\pi^+(a|s,0) \equiv \lim_{y^*\to-\infty}\left\{\int_{y^*}^{\infty} 1-\Phi_a^-(y|s,0)\,\mathrm{d}y + y^*\right\}. \tag{17}$$

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 (15), (16) and (17) into Equation (14) and simplifying the terms yield the sharp bounds on $\tau(a,0|0)$,

$$\begin{aligned}
&\sum_{s\in A}\left\{\pi^-(a|s,0)\Pr(S_i=s|A_i=0,D_i=0)\right\}-\mathbb{E}[Y_i|A_i=0,D_i=0]\\
&\qquad\qquad\qquad\leq\ \tau(a,0|0)\ \leq \qquad\qquad\qquad (18)\\
&\qquad\qquad\sum_{s\in A}\left\{\pi^+(a|s,0)\Pr(S_i=s|A_i=0,D_i=0)\right\}-\mathbb{E}[Y_i|A_i=0,D_i=0]
\end{aligned}$$

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

$$\begin{aligned}
&\sum_{s\in A}\left\{\pi^-(a|s,0)-\pi^+(a'|s,0)\right\}\Pr(S_i=s|A_i=0,D_i=0)\\
&\qquad\qquad\qquad\leq\ \tau(a,a'|0)\ \leq \qquad\qquad\qquad (19)\\
&\qquad\qquad\sum_{s\in A}\left\{\pi^+(a|s,0)-\pi^-(a'|s,0)\right\}\Pr(S_i=s|A_i=0,D_i=0)
\end{aligned}$$

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 A.2 implies that (18) and (19) 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. ▮

Proof of Proposition 2

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

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

$$\times \Pr(S_i = s, C_i = c, Y_i(0) = y_0, Y_i(1) = y_1, Y_i(2) = y_2) \Pr(D_i = d), \tag{20}$$

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 (20), $\Pr(S_i = s, C_i = c, Y_i(0) = y_0, Y_i(1) = y_1, Y_i(2) = y_2)$, 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 $\sum_{y0\in\{0,1\}} \sum_{y1\in\{0,1\}} \sum_{y2\in\{0,1\}} \sum_{s'\in\mathcal{A}} \sum_{c'\in\mathcal{A}} \phi_{y_a,y_{a'},y_{a''},s',c'} = 1$. 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):

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

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

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 one of them being redundant:

$$\Pr(S_i = s \mid A_i = a, D_i = 1) \;=\; \Pr(S_i = s) \;=\; \sum_{a \in \mathcal{A}} \sum_{y_a \in \{0,1\}} \sum_{c \in \mathcal{A}} \phi_{y_0,y_1,y_2,s,c}, \tag{23}$$

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

for all s and $a \in \mathcal{A}$. However, note that equation (23) are merely linear combinations of equation (21) 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,

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

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. ∎

Uncertainty Estimation 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. To do this, 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,

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

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($\mathbf{0}$) and $\boldsymbol{q}_s \sim$ Dirichlet($\mathbf{0}$) $\forall\ s \in \mathcal{A}$. 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:

$$\begin{bmatrix} \pi^-(a|s,c) \\ \pi^+(a|s,c) \end{bmatrix} \mid \boldsymbol{q}, \boldsymbol{X} \xrightarrow{d} \text{Normal}\left(\begin{bmatrix} \bar{\pi}^-(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) \\ \bar{\pi}^+(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) \end{bmatrix}, \begin{bmatrix} V^-(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) & C(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) \\ C(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) & V^+(a|s,c,\boldsymbol{q}_s,\boldsymbol{X}) \end{bmatrix} \right), \tag{24}$$

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 $P(a|s, c)$ (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 5 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}\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_{\hat{\Gamma}_a^{-1}(1-b)}^{\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 \\
&\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}\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 \\
&= \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 (24) 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 estimators of these parameters provided above reduce to the sample mean and the sampling variance for the mean, respectively, for the corresponding subgroup:

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

and

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

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

Uncertainty Estimation for the Sensitivity Analysis

Our approach to statistical inference for the sensitivity analysis in Section 6 is similar to the procedure outlined in Section 5. 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, ρ, 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 calculate $\boldsymbol{\tau}^-_\rho$ and $\boldsymbol{\tau}^+_\rho$ by solving the linear programming problems

$$\tau^-_\rho(a, a'|c)^* \;=\; \min_{\Pi} \quad \sum_{s\in\mathcal{A}} \{\pi(a|s, c) - \pi(a'|s, c)\} \Pr(S_i = s|A_i = c, D_i = 0) \text{ and}$$

$$\tau^+_\rho(a, a'|c)^* \;=\; \max_{\Pi} \quad \sum_{s\in\mathcal{A}} \{\pi(a|s, c) - \pi(a'|s, c)\} \Pr(S_i = s|A_i = c, D_i = 0).$$

s.t.

$$\pi^-(a^*|s, c) \;\leq\; \pi(a^*|s, c) \;\leq\; \pi^+(a^*|s, c) \quad \text{and}$$

$$\eta(a^*|c) - \rho \leq \sum_{s\in\mathcal{A}} \pi(a^*|s, c) \Pr(S_i = s|A_i = c, D_i = 0) \leq \eta(a^*|c) + \rho$$

for $a^* \in \{a, a'\}$ and $s \in \mathcal{A}$, where $\Pi \equiv \{\pi(a^*|s, c) : a^* \in \{a, a'\}, s \in \mathcal{A}\}$.

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

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

which 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

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

$$\eta(a|s) \;\perp\!\!\!\perp\; \pi^+(a'|s, c), \pi^-(a'|s, c), \eta(a'|s) \mid \hat{\boldsymbol{F}}^1, \hat{\boldsymbol{F}}^0, \boldsymbol{q}$$

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 asymptotically given by

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

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

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

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

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.

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

$$+\frac{1}{n_{sa}^{1}\cdot q_{sc}}\int_{\underline{y}}^{\hat{\Gamma}_a^{-1}(b)}\int_{\hat{\Gamma}_a^{-1}(b)}^{\infty}F(y|s,a,1)\left(1-F(x|s,a,1)\right)\,\mathrm{d}x\mathrm{d}y$$

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

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 $\boldsymbol{p} \equiv [p_s] \sim$ 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 $\boldsymbol{n}_s^0 \equiv [n_{sa}^0] = \Big[\sum_{i=1}^{N}\mathbf{1}\{S_i = s, A_i = 0, D_i = 0\}, \cdots, \sum_{i=1}^{N}\mathbf{1}\{S_i = s, A_i = J-1, D_i = 0\}\Big]^{\top}$;

   (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 $[\tau_{\rho}^{-}(a,a'|c), \tau_{\rho}^{+}(a,a'|c)]$ by solving the linear programming problems at the beginning of this appendix.

References

Arceneaux, Kevin, Martin Johnson and Chad Murphy. 2012. "Polarized Political Communication, Oppositional Media Hostility, and Selective Exposure." Journal of Politics 74(1):174–186.

Balke, Alexander. 1995. Probabilistic Counterfactuals: Semantics, Computation, and Applications PhD thesis Computer Science Department, University of California, Los Angeles.

Balke, Alexander and Judea Pearl. 1997. "Bounds on treatment effects from studies with imperfect compliance." Journal of the American Statistical Association 92:1171–1176.

Brown, Norman R and Robert C Sinclair. 1999. "Estimating Number of Lifetime Sexual Partners: Men and Women Do It Differently." Journal of Sex Research 36:292—297.

Campbell, Angus, Philip E Converse, Warren Miller and Donald Stokes. 1960. The American Voter. Chicago: University of Chicago Press.

Clausen, Aage R. 1968. "Response Validity: Vote Report." Public Opinion Quarterly 32(4):588–606.

Frangakis, Constantine E. and Donald B. Rubin. 2002. "Principal Stratification in Causal Inference." Biometrics 58(1):21–29.

Gaines, Brian J. and James H. Kuklinski. 2011. "Experimental Estimation of Heterogeneous Treatment Effects Related to Self-Selection." American Journal of Political Science 55(3):724–736.

Gallup. 2014. Media Use and Evaluation. Technical report Gallup Historical Trends.
URL: *http://www.gallup.com/poll/1663/media-use-evaluation.aspx*

Hamilton, James T. 2005. The Market and the Media. In The Press, ed. Geneva Overholser and Kathleen H Jamieson. Oxford: Oxford University Press.

Hirano, Keisuke, Guido W Imbens, Donald B Rubin and Xiao-Hua Zhou. 2000. "Assessing the effect of an influenza vaccine in an encouragement design." Biostatistics 1(1):69–88.

Horowitz, Joel L. 2001. The Bootstrap. In Handbook of Econometrics. Oxford University Press chapter 52.

Horowitz, Joel L. and Charles F. Manski. 2000. "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data." Journal of the Americal Statistical Association 95(449):77–84.

Howard, Louise and Graham Thornicroft. 2006. "Patient preference randomised controlled trials in mental health research." The British Journal of Psychiatry 188(4):303–304.

Hser, Yih-ing, Margaret Maglione and Kathleen Boyle. 1999. "Validity of Self-Report of Drug Use Among STD Patients, ER Patients, and Arrestees." American Journal of Drug and Alcohol Abuse 25(1):81–91.

Imai, Kosuke, Dustin Tingley and Teppei Yamamoto. 2013. "Experimental Designs for Identifying Causal Mechanisms (with discussions)." Journal of the Royal Statistical Society, Series A (Statistics in Society) 176(1):5–51.

Iyengar, Shanto and Kyu S. Hahn. 2009. "Red media, blue media: evidence of ideological selectivity in media use." Journal of Communication 59:19–39.

Kim, Young Mie. 2009. "Issue Publics in the New Information Environment: Selectivity, Domain Specificity, and Extremity." Communication Research 36:254–284.

King, Michael, Irwin Nazareth, Fiona Lampe, Peter Bower, Martin Chandler, Maria Morou, Bonnie Sibbald and Rosalind Lai. 2005. "Impact of participant and physician intervention preferences on randomized trials: a systematic review." Journal of the American Medical Association 293(9):1089–1099.

Ladd, Jonathan M. 2012. Why Americans Hate the Media and How It Matters. Princeton: Princeton University Press.

Levendusky, Matthew S. 2013. "Why Do Partisan Media Polarize Viewers?" American Journal of Political Science 57(3):611–623.

Long, Qi, Roderick J Little and Xihong Lin. 2008. "Causal inference in hybrid intervention trials involving treatment choice." Journal of the American Statistical Association 103(482):474–484.

Manski, Charles F. 1995. Identification Problems in the Social Sciences. Harvard University Press.

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.

Payne, Gregory J. 2010. "The Bradley Effect: Mediated Reality of Race and Politics in the 2008 U.S. Presidential Election." American Behavior Scientist 54:417–435.

Prior, Markus. 2007. Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections. Cambridge: Cambridge University Press.

Prior, Markus. 2009. "The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure." Public Opinion Quarterly 73:1–14.

Rogers, Todd and Masa Aida. 2013. "Vote Self-Prediction Hardly Predicts Who Will Vote, And Is (Misleadingly) Unbiased." American Politics Research 42(3):503–528.

Rosenbaum, Paul R. 2002. Observational Studies. 2nd ed. New York: Springer-Verlag.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." Journal of Educational Psychology 66(5):688–701.

Rubin, Donald B. 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.

Stroud, Natalie J. 2011. The Politics of News Choice. Oxford: Oxford University Press.

Tourangeau, Roger. 1999. Remember What Happened: Memory Errors and Survey Reports. In Memory: The Science of Self Report: Implications for Research and Practice, ed. Arthur A Stone, Jaylan S Turkkan, Christine A Bachrach, Jared B Jobe, Howard S Kurtzman and Virginia S Cain. Hove: Psychology Press.

Yamamoto, Teppei. 2012. "Understanding the past: Statistical analysis of causal attribution." American Journal of Political Science 56(1):237–256.