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: October 14, 2014

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 for the first time 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 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 David Nickerson and participants at the 2014 Society for Political methodology Summer Meeting for their helpful comments and suggestions.

†Ph.D. Student, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: dcknox@mit.edu.

‡Assistant 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.

1 Introduction

Randomized controlled trials (RCTs) are widely used in the social and medical sciences to estimate the causal effects of treatments of interest. The random assignment of treatments ensures the internal validity of the study, in the sense that observed differences in the distribution of outcomes between randomized treatment groups can be interpreted as causal effects of the treatments. Carefully controlled randomization, however, often comes at the cost of external validity. That is, conclusions from RCTs may not generalize to 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 bounds on this causal quantity for alternative types of outcomes, including the sharp bounds on binary outcomes. 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 investigated how the effect of partisan political news media on the subjects' perception about media and political behavior varies depending on whether they would actually consume such partisan media if they could choose.

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 5 discuss the proposed methodology. Section 6 applies the method to the empirical example. Section 7 concludes.

2 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 in-

creased 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 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 are 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, seemed 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 apparent gap between self-reported preferences or prior behavior and actual behavior, like media consumption, are manifold. These theories range from a bias toward offering socially desirable responses on topics like voting (Clausen, 1968) 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 6.

3 Design and Assumptions

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

identification analysis.

3.1 Notation and the Proposed Design

Suppose that we have a random sample of N experimental subjects from the population of interest. We consider a study where the goal is to estimate the effect of a J-valued treatment on an outcome of interest. Let $A_i \in \mathcal{A} \equiv \{0, 1, ..., J-1\}$ denote the treatment that subject i actually receives in the study. For the rest of the paper, we call this interchangeably the "actual treatment," "observed choice," or simply the "treatment" when the meaning is obvious from the context.

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. This 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. Below, we consider three types of outcomes, with decreasing generality: unbounded ($\mathcal{Y} \subseteq \mathbb{R}$), bounded ($\mathcal{Y} = [\underline{y}, \overline{y}]$), and binary ($\mathcal{Y} = \{0, 1\}$). 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 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 6. Finally, note that we allow the treatment variable to be multi-valued, instead of binary. In fact, as previously shown by 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 causal 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 prefernces and another

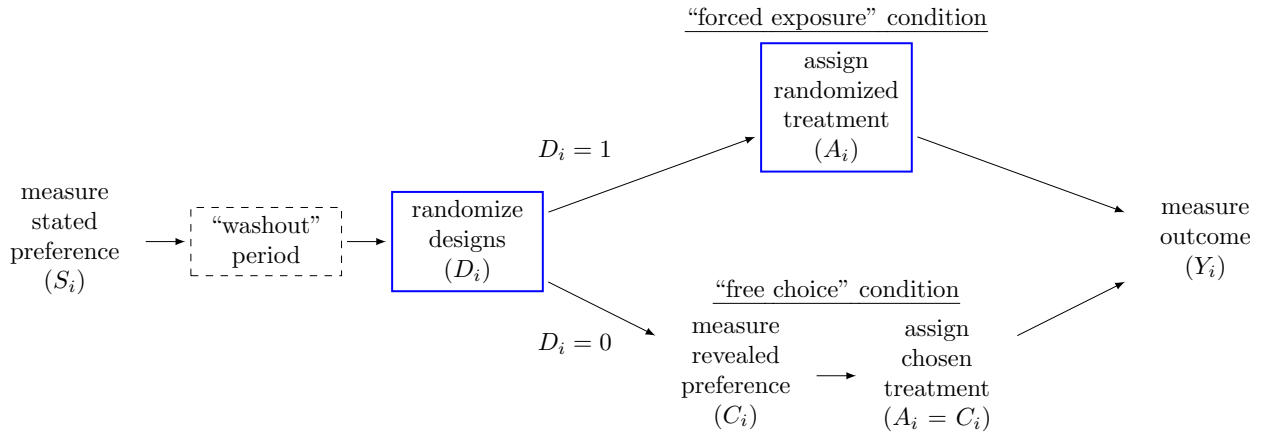

Figure 1: Diagram of the Proposed PPT Design. In the proposed design, subjects are first asked to state preferences about the treatment options (S_i) and (after an optional "washout" period) randomized into design conditions (D_i). In the "forced exposure" arm (top, $D_i = 1$), subjects are randomly 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.

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.

3.2 Quantities of interest

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

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

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

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

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., $\delta(a, a') = \sum_c \tau(a, a' \mid c) \Pr(C_i = c)$).

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

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

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.

3.3 Assumptions

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

Assumption 1 (Randomization of Designs)

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

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)

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

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.

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

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.

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

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 variances, obtained via standard techniques like the delta method.

4 Nonparametric Identification Analysis

In this section, we present the results of our nonparametric identification analysis for the ACTE under the proposed design. We begin with the results that hold for the most general class of outcomes ($\mathcal{Y} \subseteq \mathbb{R}$) and then consider more restricted set of outcomes. We ends our analysis with the derivation of the sharp (i.e., tightest possible) nonparametric bounds for binary outcomes.

4.1 General Results for Unbounded Outcomes

We first present our most general results that are valid for any real-valued outcome ($\mathcal{Y} \subseteq \mathbb{R}$). 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) and the SUTVA. In Appendix A.2, we show that the ACTE can be expressed as follows under those assumptions.

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

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.

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

for a, a' and $c \in \{0, 1\}$. This exactly matches Gaines and Kuklinski's (2011, p.729), 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 $J = 2$. Thus, we verify their earlier result under the potential outcomes 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 $\mathbb{E}[Y_i(a'') \mid C_i = c'] = \mathbb{E}[Y_i \mid A_i = a'', S_i = c', D_i = 1]$ 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.

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

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? Unfortunately, equation (4) does not imply any restriction on the possible value of $\tau(a, a' \mid c)$. Here, we take an alternative approach and obtain the following partial identification result for the general case of unbounded outcomes.

Proposition 1 (Nonparametric Bounds on the ACTE for Unbounded Outcomes) *Under Assumptions 1 and 2, the ACTE can be partially identified at least up to the following nonparametric bounds:*

$$\begin{aligned} \mathbb{E}[Y_i \mid A_i = a, D_i = 1] - \mathbb{E}[Y_i \mid A_i = a', D_i = 1] + \frac{\max_{s \in \mathcal{A}} \{Q(a) - R(a)\} - \min_{s \in \mathcal{A}} \{Q(a') + R(a')\}}{\Pr(C^*_{ic} = 1 \mid D_i = 0)} \\ \leq \; \tau(a, a' \mid c) \; \leq \\ \mathbb{E}[Y_i \mid A_i = a, D_i = 1] - \mathbb{E}[Y_i \mid A_i = a', D_i = 1] + \frac{\min_{s \in \mathcal{A}} \{Q(a) + R(a)\} - \max_{s \in \mathcal{A}} \{Q(a') - R(a')\}}{\Pr(C^*_{ic} = 1 \mid D_i = 0)} \end{aligned} \tag{6}$$

for any a, a' and $c \in \mathcal{A}$, where

$$
\begin{aligned}
Q(a^*) &= \{\mathbb{E}[Y_i \mid S^*_{is} = 1, A_i = a^*, D_i = 1] - \mathbb{E}[Y_i \mid S^*_{is} = 0, A_i = a^*, D_i = 1]\} \operatorname{Cov}(S^*_{is}, C^*_{ic} \mid D_i = 0), \\
R(a^*) &= \sqrt{\operatorname{Var}(Y_i \mid A_i = a^*, D_i = 1)} \sqrt{\operatorname{Var}(C^*_{ic} \mid D_i = 0)} \\
&\qquad \times \sqrt{1 - \operatorname{Cor}(Y_i, S^*_{is} \mid A_i = a^*, D_i = 1)^2} \sqrt{1 - \operatorname{Cor}(S^*_{is}, C^*_{ic} \mid D_i = 0)^2},
\end{aligned}
$$

for $a^* \in \{a, a'\}$, $C^*_{ic} = \mathbf{1}\{C_i = c\}$, and $S^*_{is} = \mathbf{1}\{S_i = s\}$.

A proof can be found in Appendix A.3. The bounds in equation (6) provide several useful insights. First, the first two terms in both the upper and lower bounds equal the ATE ($\delta(a, a')$), which represents the overall average effect of exposing units to treatment a as opposed to a' regardless of their preferences. Second, $Q(a)$ and $Q(a')$ can be considered adjustment terms for the ATE which incorporate the selection effect by use of the stated preference information. These terms equal zero under either following conditions: (1) the potential outcomes under treatments a and a' are on average equal between those who state they prefer to take treatment s and those who prefer not to, or (2) the stated and revealed preferences are uncorrelated. Neither condition is likely to hold exactly in situations where the ACTE becomes an object of the study. Third, the expression for $R(a)$ and $R(a')$ implies that the bounds become tighter under several alternative conditions: (1) when the potential outcomes are less variable, (2) when the proportion of subjects who choose the treatment in consideration is close to half of the population, (3) when the potential outcomes are highly correlated with the stated choice, and (4) when the stated and revealed preferences have a high correlation. As anticipated, the bounds have zero length, and the ACTE is therefore point identified as equation (5), when stated and revealed preferences are perfectly correlated.

Finally, note that when the subgroup of interest would choose $C_i = a'$ (or a), the bounds in equation (6) can be simplified and tightened. This is because the corresponding conditional average potential outcome is point identified from the free choice arm, i.e., $\mathbb{E}[Y_i(c) \mid C_i = c] = \mathbb{E}[Y_i \mid A_i = c, D_i = 0]$. The simplified bounds for this imporant special case are given in Appendix A.3.

4.2 Bounds on the ACTE for Bounded Outcomes

The nonparametric bounds on the ACTE in Section 4.1 can be further tightened when we focus on finitely bounded outcomes ($\mathcal{Y} = [\underline{y}, \overline{y}]$). The following proposition describes the result.

Proposition 2 (Nonparametric Bounds on the ACTE for Bounded Outcomes) *Under Assumptions 1 and 2, the ACTE for a bounded outcome is guaranteed to exist within the intersection of the interval given in Proposition 1 and the interval given by the following upper and lower bounds.*

$$\begin{aligned}
&\sum_{s=0}^{J-1} \{\underline{\pi}(a \mid s, c) - \overline{\pi}(a' \mid s, c)\} \Pr(S_i = s \mid C_i = c, D_i = 0) \\
&\qquad \leq \ \tau(a, a' \mid c) \ \leq \ \sum_{s=0}^{J-1} \{\overline{\pi}(a \mid s, c) - \underline{\pi}(a' \mid s, c)\} \Pr(S_i = s \mid C_i = c, D_i = 0), \qquad (7)
\end{aligned}$$

where

$$\underline{\pi}(a^* \mid s, c) \ = \ \begin{cases} \max \left[\underline{y}, \ \frac{1}{\Pr(C_i = c \mid S_i = s, D_i = 0)} \left\{ \begin{array}{l} \mathbb{E}[Y_i \mid S_i = s, A_i = a^*, D_i = 1] \\ -\mathbb{E}[Y_i \mid S_i = s, C_i = a^*, D_i = 0] \\ \qquad \times \Pr(C_i = a^* \mid S_i = s, D_i = 0) \\ -\sum_{c' \notin \{a^*, c\}} \overline{y} \Pr(C_i = c' \mid S_i = s, D_i = 0) \end{array} \right\} \right] & \text{if } a^* \neq c, \\ \mathbb{E}[Y_i \mid S_i = s, C_i = c, D_i = 0] & \text{if } a^* = c, \end{cases}$$

$$\overline{\pi}(a^* \mid s, c) \ = \ \begin{cases} \min \left[\overline{y}, \ \frac{1}{\Pr(C_i = c \mid S_i = s, D_i = 0)} \left\{ \begin{array}{l} \mathbb{E}[Y_i \mid S_i = s, A_i = a^*, D_i = 1] \\ -\mathbb{E}[Y_i \mid S_i = s, C_i = a^*, D_i = 0] \\ \qquad \times \Pr(C_i = a^* \mid S_i = s, D_i = 0) \\ -\sum_{c' \notin \{a^*, c\}} \underline{y} \Pr(C_i = c' \mid S_i = s, D_i = 0) \end{array} \right\} \right] & \text{if } a^* \neq c, \\ \mathbb{E}[Y_i \mid S_i = s, C_i = c, D_i = 0] & \text{if } a^* = c, \end{cases}$$

where $a^* \in \{a, a'\}$*, for all* a, a' *and* $c \in \mathcal{A}$*.*

A proof is given in Appendix A.4. We offer several remarks about Proposition 2. First, these bounds are naturally more informative when more units choose one of the treatments of interest, as worst-case assumptions then apply to a smaller portion of the population. Second, bounds are also narrower for treatments that are more likely to be chosen. Third, bounds tend to be more informative when the general population responds differently than units that self-select into a treatment of interest. Finally, as before, we can generally obtain tighter bounds for the ACTE where one of the two potential outcomes

refers to the condition actually observed in the real world (i.e., $a = c$ or $a' = c$). This is again because we can point-identify the conditional mean of the potential outcome corresponding to that condition using the free choice group. For example, the upper and lower bounds on $\tau(a, c \mid c)$ are given by

$$\sum_{s=0}^{J-1} \underline{\pi}(a \mid s, c) \Pr(S_i = s | C_i = c, D_i = 0) - \mathbb{E}[Y_i \mid C_i = c, D_i = 0]$$
$$\leq \tau(a, c \mid c) \leq \sum_{s=0}^{J-1} \overline{\pi}(a \mid s, c) \Pr(S_i = s | C_i = c, D_i = 0) - \mathbb{E}[Y_i \mid C_i = c, D_i = 0]$$

for any a and $c \in \mathcal{A}$. Additional details are also provided in Appendix A.4.

4.3 Sharp Bounds on the ACTE for Binary Outcomes

We now further restrict analysis to binary outcomes ($\mathcal{Y} = \{0, 1\}$) and derive another set of nonparametric bounds for the ACTE. In this case, we can in fact obtain the *sharp* bounds (i.e., the tightest possible given all the observed information; Manski, 1995) by incorporating the full joint distribution of the observed variables in the derivation of the bounds. This implies that the resulting bounds cannot be improved without introducing additional assumptions that are not justified by the experimental design itself.

Specifically, we take 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 $\phi_{y_0,y_1,y_2,s,c}$. Also, note that 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 3 (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.5. The maximization and minimization problems in Proposition 3 are standard linear programming problems which can be easily solved numerically with given data using statistical software, such as the `lpSolve` package in R.

5 Sensitivity Analysis

The nonparametric bounds we have derived so far all 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,

$$\rho \;\equiv\; \max \Big| \mathbb{E}[Y_i(a) \mid S_i = c] - \mathbb{E}[Y_i(a) \mid C_i = c] \Big|$$

for all a and $c \in \mathcal{A}$. In words, ρ represents the maximum absolute difference we allow to exist between the average potential outcome among those who state to choose a particular treatment and the average potential outcome among those who actually choose that treatment. This definition of ρ implies the following additional constraint,

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

for all a and $s \in \mathcal{A}$.

The proposed sensitivity analysis proceeds by incorporating equation (8) to the calculation of bounds. Specifically, for the case of unbounded outcomes, we modify the bounds given in Proposition 1 to

$$\begin{aligned}
&\mathbb{E}[Y_i \mid A_i = a, D_i = 1] - \mathbb{E}[Y_i \mid A_i = a', D_i = 1] \\
&\qquad + \max \left[\left\{ \frac{Q(a) - R(a)}{\Pr(C_{ic}^* = 1 \mid D_i = 0)} : s \in \mathcal{A} \right\} \cup \{\mathbb{E}[Y_i \mid S_i = c, A_i = a, D_i = 1] - \rho\} \right] \\
&\qquad - \min \left[\left\{ \frac{Q(a') + R(a')}{\Pr(C_{ic}^* = 1 \mid D_i = 0)} : s \in \mathcal{A} \right\} \cup \{\mathbb{E}[Y_i \mid S_i = c, A_i = a, D_i = 1] + \rho\} \right] \\
&\qquad\qquad \leq\; \tau(a, a' \mid c) \;\leq \\
&\mathbb{E}[Y_i \mid A_i = a, D_i = 1] - \mathbb{E}[Y_i \mid A_i = a', D_i = 1] \\
&\qquad + \max \left[\left\{ \frac{Q(a) + R(a)}{\Pr(C_{ic}^* = 1 \mid D_i = 0)} : s \in \mathcal{A} \right\} \cup \{\mathbb{E}[Y_i \mid S_i = c, A_i = a, D_i = 1] + \rho\} \right] \\
&\qquad - \min \left[\left\{ \frac{Q(a') - R(a')}{\Pr(C_{ic}^* = 1 \mid D_i = 0)} : s \in \mathcal{A} \right\} \cup \{\mathbb{E}[Y_i \mid S_i = c, A_i = a, D_i = 1] - \rho\} \right]
\end{aligned}$$

for all a, a' and $c \in \mathcal{A}$.

For bounded 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 numerically for a given ρ by solving the following linear programming problem.

$$\min_{\Pi_{a,a'}} \quad \text{and} \quad \max_{\Pi_{a,a'}} \quad \sum_{s' \in \mathcal{A}} \pi(a^* \mid s', c) \Pr(S_i = s' \mid C_i = c, D_i = 0),$$

s.t. $\pi(a^* \mid s', c') \geq \underline{y} \;\forall\; a^*, s', c'$, $\pi(a^* \mid s', c') \leq \overline{y} \;\forall\; a^*, s', c'$, $\pi(a^* \mid s', a^*) = \mathbb{E}[Y_i \mid S_i = s', C_i =$

$a^*, D_i = 0] \; \forall \; a^*, \sum_{c' \in \mathcal{A}} \pi(a^* \mid s', c') \Pr(C_i = c' \mid S_i = s', D_i = 0) = \mathbb{E}[Y_i \mid S_i = s', A_i = a^*, D_i = 1] \; \forall \; a^*, s', \sum_{s' \in \mathcal{A}} \pi(a^* \mid s', c') \Pr(S_i = s' \mid C_i = c', D_i = 0) \geq \mathbb{E}[Y_i \mid S_i = c', A_i = a^*, D_i = 1] - \rho \; \forall \; a^*, c', \sum_{s' \in \mathcal{A}} \pi(a^* \mid s', c') \Pr(S_i = s' \mid C_i = c', D_i = 0) \leq \mathbb{E}[Y_i \mid S_i = c', A_i = a^*, D_i = 1] + \rho \; \forall \; a^*, c'$, where $\Pi_{a,a'} \equiv \{\pi(a^* \mid s', c) : a^* \in \{a, a'\}, s \in \mathcal{A}, c \in \mathcal{A}\}$ and $\pi(a \mid s, c) \equiv \mathbb{E}[Y_i(a) \mid S_i = s, C_i = c]$. The first and second constraints represent the range of the outcome, the third and fourth respectively incorporate observed outcomes from the free-choice and forced-exposure arms, and the fifth and sixth together impose equation (8).

Finally, for binary outcomes, we incorporate equation (8) into the linear programming problem in Proposition 3 as another set of linear constraints. For the special case of $J = 3$, these additional 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 $c', a'' \in \mathcal{A}$.

6 Empirical Application

In this section, we apply the proposed methodology to the empirical example we described in Section 2.

6.1 Design and Data

In implementing the media choice experiment, we closely followed the proposed protocol as described in Section 3.1 and summarized in Figure 1. First, to measure the stated preferences over treatment options, 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 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 actually choices, and that those stating a preference for entertainment were significantly more likely to be consistent in their actual choices ($\Pr(C_i = 0 \mid S_i = 0) = 0.91$, whereas $\Pr(C_i = 1 \mid S_i = 1) = 0.81$ and $\Pr(C_i = -1 \mid S_i = -1) = 0.77$; a χ^2 test of independence between stated preference and consistency has a p-value on the order of 10^{-6}). 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

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.

are also associated with different responses to media. For example, among free-choice units that stated a preference for pro-attitudinal media and also chose it, 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.

6.2 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 (bounded continuous; Proposition 2), and the right panel presents results for whether respondents were likely to discuss the story with a friend (binary; Proposition 3). Each vertically arrayed plot depicts 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 leftmost 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

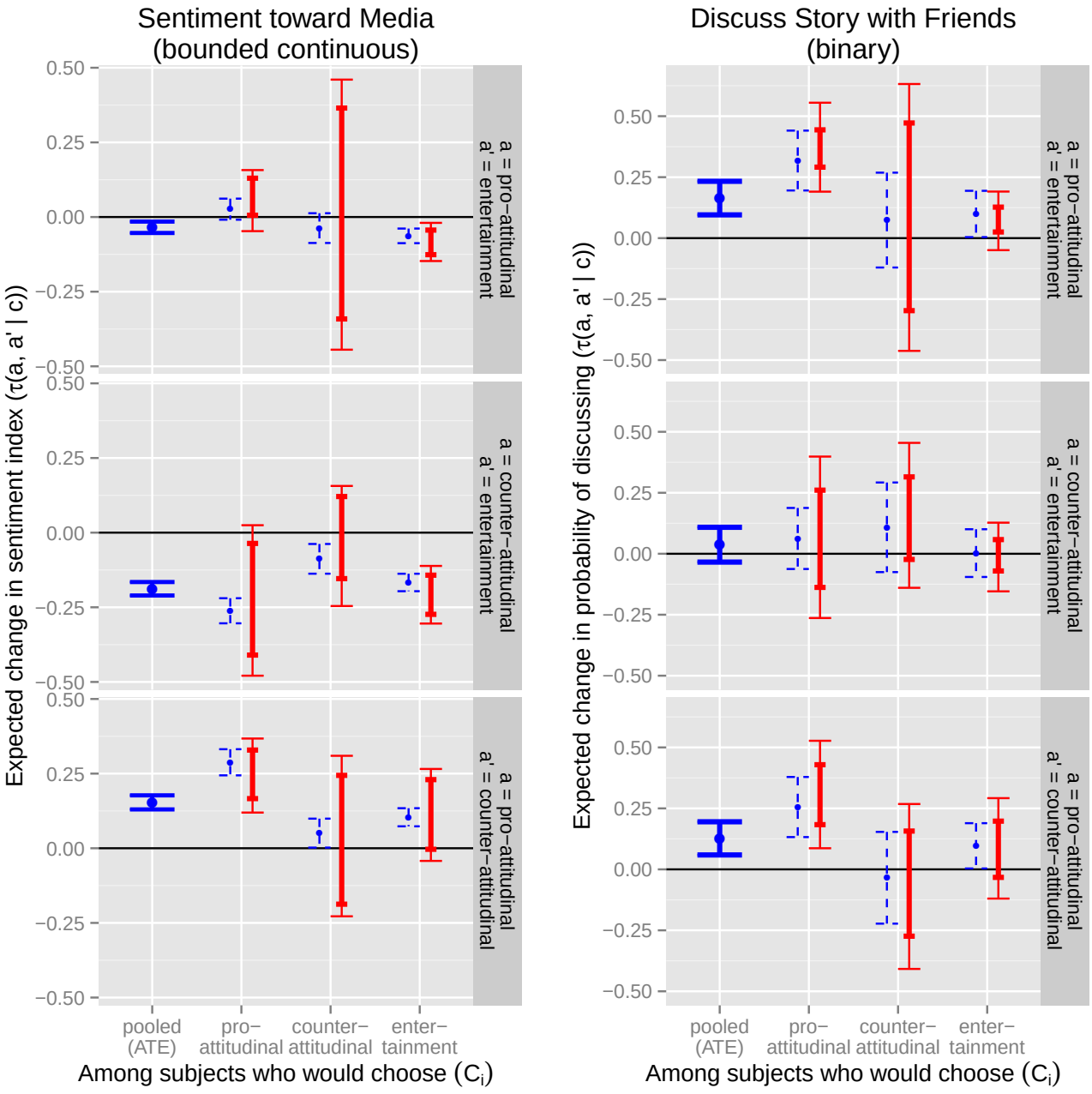

Figure 2: Estimated Nonparametric Bounds on the ACTE of Partisan News Media. Vertically stacked plots correspond to the same outcome variable. Horizontally aligned plots depict the effect of a particular change in the assigned media, i.e., $\mathbb{E}[Y_i(a) - Y_i(a') \mid C_i = c]$. Pairs of lines correspond 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.

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 6.3, 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 have substantially higher affect 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 4.3, 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).

6.3 Sensitivity Analysis

Next, we apply the sensitivity analysis developed in Section 5 and show how the bounds become tighter as we allow less difference between the average potential outcomes conditional on a stated preference versus actual choice (ρ). For illustration, we focus on the 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.

7 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, sysmatic 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

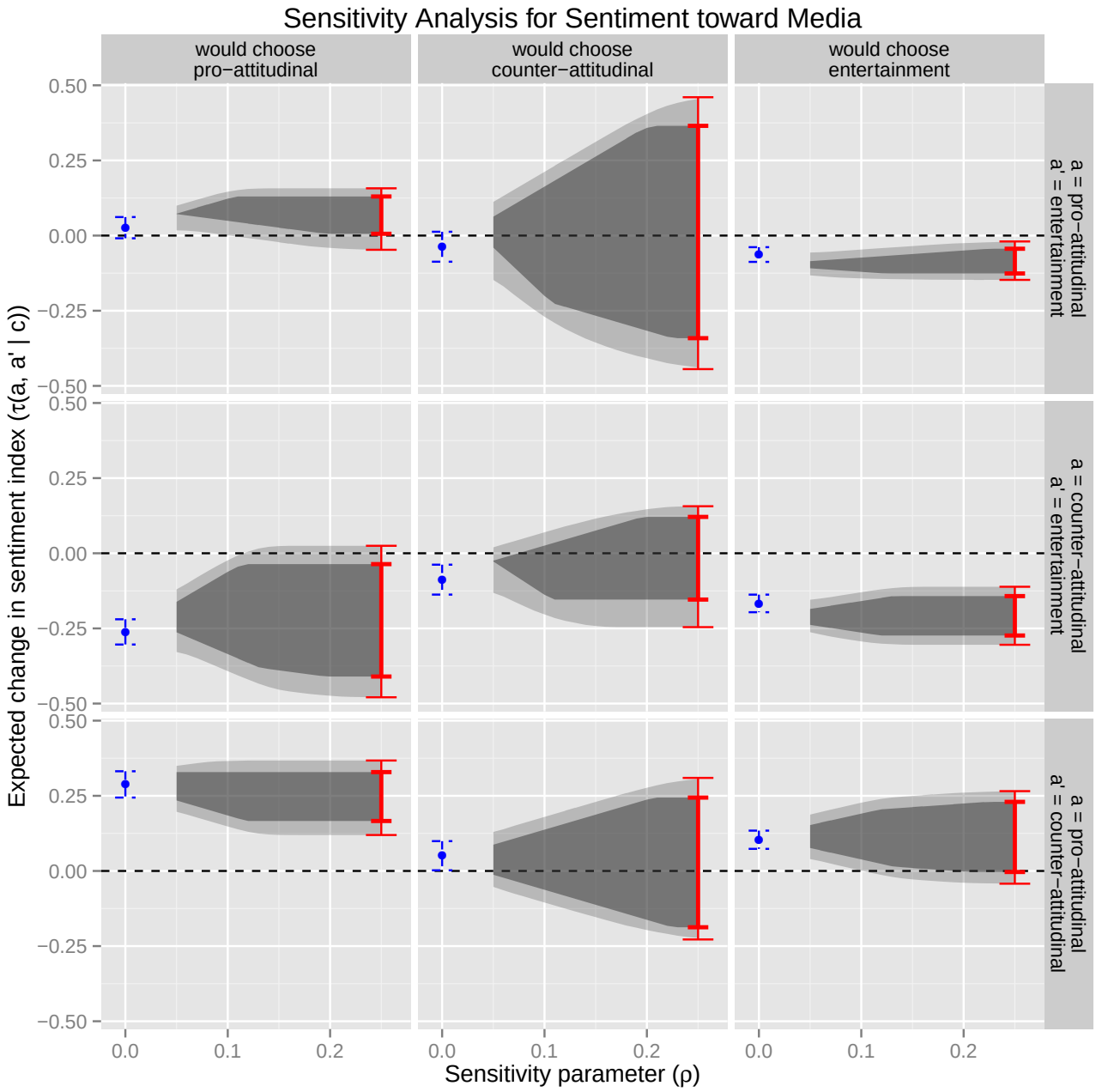

Figure 3: Sensitivity Analysis for the ACTE of Partisan News Media. The plots correspond to the left panel of Figure 2. On the left side of each plot, a blue point and error bars represent the naïve estimate and 95% asymptotic confidence intervals, respectively. On the right side, thick red error bars 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.

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.

A Appendix

A.1 Observable Implications of Assumption 3

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

$$\mathbb{E}[Y_i(a) \mid C_i = a] \;=\; \mathbb{E}[Y_i(a) \mid S_i = a], \tag{9}$$

for all $a \in \mathcal{A}$. This relationship directly implies equation (2) under Assumptions 1 and 2. Second, note that equation (9) also implies,

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

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

A.2 Derivation of Equation (4)

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

1], and $\mathbb{E}[Y_i(a) \mid S_i = s] = \mathbb{E}[Y_i \mid S_i = s, A_i = a, D_i = 1]$. Now, note that

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

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

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

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

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

The quantity of interest is therefore

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

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

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

and $J-2$ terms remain unidentified.

A.3 Proof of Proposition 1

We proceed by considering $\mathbb{E}[Y_i(a) \mid C_i = c]$ first. This quantity is related to C_{ic}^* by

$$\begin{aligned}\mathrm{Cor}(Y_i(a), C_{ic}^*) &= \frac{\mathbb{E}[Y_i(a)C_{ic}^*] - \mathbb{E}[Y_i(a)]\mathbb{E}[C_{ic}^*]}{\sqrt{\mathrm{Var}(Y_i(a))\mathrm{Var}(C_{ic}^*)}} \\ &= \sqrt{\frac{\Pr(C_i = c)}{\sigma_a^2\{1 - \Pr(C_i = c)\}}}\left(\mathbb{E}[Y_i(a) \mid C_i = c] - \mu_a\right)\end{aligned}$$

where $\mu_a = \mathbb{E}[Y_i \mid A_i = a, D_i = 1]$, and $\sigma_a^2 = \mathrm{Var}(Y_i \mid A_i = a, D_i = 1)$. The last equality holds because of Assumptions 1, 2 and the law of total expectation. Now, note that for any three random variables X, W, and Z, the Cauchy-Schwarz inequality implies

$$\mathrm{Cor}(X, W) \in \left[\begin{array}{c} \mathrm{Cor}(X,Z)\mathrm{Cor}(W,Z) - \sqrt{1 - \mathrm{Cor}(X,Z)^2}\sqrt{1 - \mathrm{Cor}(W,Z)^2}, \\ \mathrm{Cor}(X,Z)\mathrm{Cor}(W,Z) + \sqrt{1 - \mathrm{Cor}(X,Z)^2}\sqrt{1 - \mathrm{Cor}(W,Z)^2} \end{array}\right].$$

Setting $X = Y_i(a)$, $W = C_{ic}^*$ and $Z = S_{is}^*$ yields the following inequalities,

$$\begin{aligned}&\mathrm{Cor}(Y_i(a), S_{is}^*)\mathrm{Cor}(C_{ic}^*, S_{is}^*) - \sqrt{1 - \mathrm{Cor}(Y_i(a), S_{is}^*)^2}\sqrt{1 - \mathrm{Cor}(C_{ic}^*, S_{is}^*)^2} \\ &\quad\leq \sqrt{\frac{\Pr(C_i = c)}{\sigma_a^2\{1 - \Pr(C_i = c)\}}}\left(\mathbb{E}[Y_i(a) \mid C_i = c] - \mu_a\right) \leq \\ &\qquad \mathrm{Cor}(Y_i(a), S_{is}^*)\mathrm{Cor}(C_{ic}^*, S_{is}^*) + \sqrt{1 - \mathrm{Cor}(Y_i(a), S_{is}^*)^2}\sqrt{1 - \mathrm{Cor}(C_{ic}^*, S_{is}^*)^2}\end{aligned} \tag{10}$$

for any $s \in \mathcal{A}$, where $S_{is}^* = \mathbf{1}\{S_i = s\}$. Now note that

$$\mathrm{Cor}(Y_i(a), S_{is}^*) = \sqrt{\frac{\Pr(S_i = s)}{\sigma_a^2\{1 - \Pr(S_i = s)\}}}\left(\mathbb{E}[Y_i \mid S_i = s, A_i = a, D_i = 1] - \mu_a\right) = \frac{\sqrt{\mathrm{Var}(S_{is}^*)}}{\sigma_a}\delta$$

where $\delta = \mathbb{E}[Y_i \mid S_i = s, A_i = a, D_i = 1] - \mathbb{E}[Y_i \mid S_i \neq s, A_i = a, D_i = 1]$, because of Assumptions 1, 2, and the law of total expectation. Similarly, $\mathrm{Cor}(C_{ic}^*, S_{is}^*) = \sqrt{\mathrm{Var}(S_{is}^*)/\mathrm{Var}(C_{ic}^* \mid D_i = 0)}\gamma$ where $\gamma = \Pr(C_i = c \mid S_i = s, D_i = 0) - \Pr(C_i = c \mid S_i \neq s, D_i = 0)$. Substituting these into equation (10) and rearranging terms yields the following upper and lower bounds on $\mathbb{E}[Y_i(a) \mid C_i = c]$ for given s,

$$\mu_a + \frac{\mathrm{Var}(S_{is}^*)\delta\gamma \pm \sqrt{(\sigma_a^2 - \mathrm{Var}(S_{is}^*)\delta^2)(\mathrm{Var}(C_{ic}^* \mid D_i = 0) - \mathrm{Var}(S_{is}^*)\gamma^2)}}{\Pr(C_i = c \mid D_i = 0)}$$

for all a and c. Taking the intersection of the intervals given for all $s \in \mathcal{A}$ and reexpressing some of the terms, we obtain

$$\mathbb{E}[Y_i \mid A_i = a, D_i = 1] + \frac{\max_{s\in\mathcal{A}} \{Q(a) - R(a)\}}{\Pr(C^*_{ic} = 1 \mid D_i = 0)}$$
$$\leq \mathbb{E}[Y_i(a) \mid C_i = c] \leq \mathbb{E}[Y_i \mid A_i = a, D_i = 1] + \frac{\min_{s\in\mathcal{A}} \{Q(a) + R(a)\}}{\Pr(C^*_{ic} = 1 \mid D_i = 0)},$$

where $Q(a)$ and $R(a)$ are defined in Section 4.1. Note that the set of bounds corresponding to $s = c$ will tend to dominate when stated preferences are closely related to actual choices so that $\mathrm{Cor}(C^*_{ic}, S^*_{ic})$ is large, all else equal. The lower bound on $\tau(a, a' \mid c)$ in equation (6) can then be obtained by differencing the lower bound of $\mathbb{E}[Y_i(a) \mid C_i = c]$ and the upper bound of $\mathbb{E}[Y_i(a') \mid C_i = c]$, and vice-versa for the upper bound on $\tau(a, a' \mid c)$.

When either $a = c$ or $a' = c$, bounds can be tightened because the corresponding conditional mean potential outcome is point-identified from the free-choice arm. The bounds for $\tau(a, c \mid c)$ are,

$$\mathbb{E}[Y_i \mid A_i = a, D_i = 1] - \mathbb{E}[Y_i \mid C^*_{ic} = 1, D_i = 0] + \frac{\max_{s\in\mathcal{A}} \{Q(a) - R(a)\}}{\Pr(C^*_{ic} = 1 \mid D_i = 0)}$$
$$\leq \tau(a, c \mid c) \leq \mathbb{E}[Y_i \mid A_i = a, D_i = 1] - \mathbb{E}[Y_i \mid C^*_{ic} = 1, D_i = 0] + \frac{\min_{s\in\mathcal{A}} \{Q(a) + R(a)\}}{\Pr(C^*_{ic} = 1 \mid D_i = 0)}$$

and the bounds for $\tau(c, a' \mid c)$ can be similarly obtained.

A.4 Proof of Proposition 2

We first divide subjects into J^2 strata defined by S_i and C_i. For each stratum, we consider the J average potential outcomes ("strata means"), represented as $\pi(a \mid s, c) \equiv \mathbb{E}[Y_i(a) \mid S_i = s, C_i = c]$. Then, Assumptions 1 and 2 imply

$$\mathbb{E}[Y_i \mid S_i = s, A_i = a, D_i = 1] = \sum_{c=0}^{J-1} \pi(a \mid s, c) \Pr(C_i = c | S_i = s, D_i = 0).$$

Rearranging and substituting,

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

Thus, the bounds of individual strata means are given by

$$\underline{\pi}(a \mid s,c) = \begin{cases} \max\{\underline{y}, \underline{\pi}^*(a \mid s,c)\} & \text{if} \quad a \neq c \\ \mathbb{E}[Y_i|S_i=s, C_i=c, D_i=0] & \text{if} \quad a = c \end{cases}$$

$$\overline{\pi}(a \mid s,c) = \begin{cases} \min\{\overline{y}, \overline{\pi}^*(a \mid s,c)\} & \text{if} \quad a \neq c \\ \mathbb{E}[Y_i|S_i=s, C_i=c, D_i=0] & \text{if} \quad a = c \end{cases},$$

where

$$\underline{\pi}^*(a \mid s,c) = \frac{1}{\Pr(C_i=c|S_i=s,D_i=0)} \left\{ \begin{array}{l} \mathbb{E}[Y_i \mid S_i=s, A_i=a, D_i=1] \\ -\mathbb{E}[Y_i|S_i=s, C_i=a, D_i=0]\Pr(C_i=a|S_i=s,D_i=0) \\ -\sum_{c' \notin \{a,c\}} \overline{y} \Pr(C_i=c'|S_i=s,D_i=0) \end{array} \right\},$$

$$\overline{\pi}^*(a \mid s,c) = \frac{1}{\Pr(C_i=c|S_i=s,D_i=0)} \left\{ \begin{array}{l} \mathbb{E}[Y_i \mid S_i=s, A_i=a, D_i=1] \\ -\mathbb{E}[Y_i|S_i=s, C_i=a, D_i=0]\Pr(C_i=a|S_i=s,D_i=0) \\ -\sum_{c' \notin \{a,c\}} \underline{y} \Pr(C_i=c'|S_i=s,D_i=0) \end{array} \right\},$$

by setting the unobserved strata means to either $\overline{y}$ or $\underline{y}$. The average choice-specific potential outcome is thus guaranteed to lie within $\sum_{s=0}^{J-1} \underline{\pi}(a \mid s,c) \Pr(S_i = s \mid C_i = c, D_i = 0) \leq \mathbb{E}[Y_i(a) \mid C_i = c] \leq \sum_{s=0}^{J-1} \overline{\pi}(a \mid s,c) \Pr(S_i = s \mid C_i = c, D_i = 0)$. (Note that this collapses to $\mathbb{E}[Y_i \mid C_i = c, D_i = 0]$ for $c = a$.) Taking the difference between the minimum of $\mathbb{E}[Y_i(a) \mid C_i = c]$ and the maximum of $\mathbb{E}[Y_i(a') \mid C_i = c]$, and vice versa, yields the bounds on $\tau(a, a' \mid c)$ in equation (7).

A.5 Proof of Proposition 3

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

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

$$\times \Pr(S_i = s, C_i = c, Y_i(0) = y_0, Y_i(1) = y_1, Y_i(2) = y_2) \cdot \Pr(D_i = d), \tag{11}$$

where the first equality is by Assumption 1 and the second equality is by assumption 2, noting also that Y_i is a deterministic function of $Y_i(0)$, $Y_i(1)$, $Y_i(2)$ and A_i and that C_i and D_i are sufficient for A_i.

In equation (11), all components are either deterministic relationships or are fixed by randomization, with the exception of $\Pr(S_i = s, C_i = c, Y_i(0) = y_0, Y_i(1) = y_1, Y_i(2) = y_2)$; therefore, this distribution completely specifies the model, with $|\mathcal{A}|^2 \cdot |\mathcal{Y}|^{|\mathcal{A}|} - 1 = J^2 2^J - 1$ free parameters needed to describe it. Balke (1995) (subsection 3.5) shows that bounds on counterfactual probabilities found by optimizing over such a complete model are sharp; that is, they are guaranteed to be at least as tight as bounds calculated from any partial (marginalized) model.

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

$$\Pr(S_i = s, C_i = c \mid D_i = 0) = \Pr(S_i = s, C_i = c) = \sum_{a \in \mathcal{A}} \sum_{y_a \in \{0,1\}} \phi_{y_0,y_1,y_2,s,c}, \tag{12}$$

$$\Pr(S_i = s, C_i = c, Y_i = 1 \mid D_i = 0) = \Pr(S_i = s, C_i = c, Y_i(c) = 1) = \sum_{a \neq c} \sum_{y_a \in \{0,1\}} \phi_{y_0,y_1,y_2,s,c}, \tag{13}$$

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:

$$\Pr(S_i = s \mid A_i = a, D_i = 1) = \Pr(S_i = s) = \sum_{a \in \mathcal{A}} \sum_{y_a \in \{0,1\}} \sum_{c \in \mathcal{A}} \phi_{y_0,y_1,y_2,s,c}, \tag{14}$$

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

for all s and $a \in \mathcal{A}$. However, note that equation (14) are merely linear combinations of equation (12) 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 3.

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. 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.

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.

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.