A Generalized Framework for the Estimation of Causal Moderation Effects with Randomized Treatments and Non-Randomized Moderators

Kirk Bansak

March 2018

ABSTRACT

Researchers are often interested in analyzing conditional treatment effects. One variant of this is “causal moderation,” which implies that intervention upon a third (moderator) variable would alter the treatment effect. This study presents a generalized, non-parametric framework for estimating causal moderation effects given randomized treatments and non-randomized moderators that achieves a number of goals. First, it highlights how conventional approaches do not constitute unbiased or consistent estimators of causal moderation effects. Second, it offers researchers a simple, transparent approach for estimating causal moderation effects and lays out the assumptions under which this can be performed consistently and/or without bias. Third, as part of the estimation process, it allows researchers to implement their preferred method of covariate adjustment, including parametric and non-parametric methods, or alternative identification strategies of their choosing. Fourth, it provides a set-up whereby sensitivity analysis designed for the average-treatment-effect context can be extended to the moderation context. An original application is also presented.

Kirk Bansak, Department of Political Science, Stanford University, Stanford, CA 94305-6044. Email: kbansak@stanford.edu.

I thank Jens Hainmueller, Justin Grimmer, Mike Tomz, Laura Stoker, Ken Scheve, and Andy Hall for helpful comments.
I. Introduction

Experimental researchers are often interested in analyzing conditional treatment effects. By better understanding how the magnitude or sign of a treatment effect may depend upon other variables, researchers can better explain the phenomena they are studying and determine the likelihood that a treatment will be effective if introduced to other populations in the future. Most commonly, researchers will investigate whether the effect of a treatment $T$ on an outcome $Y$ is conditional upon the value of a third variable $S$.

However, if researchers find evidence of treatment effect conditionality, they then face an additional hurdle in interpreting the nature of that conditionality. On the one hand, they may simply observe a difference in the treatment effect across different values of $S$ because the treatment is more or less impactful for different segments of the population. This phenomenon is often referred to as treatment effect heterogeneity, which is observable variation in the effect of $T$ on $Y$ across sub-populations with different observable characteristics. In this case, it can be established that the treatment effect is indeed different across sub-groups with different values of $S$, but what cannot be established is why or whether that difference is actually attributable to $S$ at all.

On the other hand, however, researchers may be interested in further claiming that $S$ is precisely the reason for the difference in treatment effects—that is, changes in $S$ actually cause changes in the treatment effect of $T$ on $Y$. This phenomenon, which is the focus of this study, will be referred to as the “causal moderation” of treatment effects, which is the causal effect of a third variable ($S$) on the treatment’s ($T$’s) effect on the outcome ($Y$). In contrast to treatment effect heterogeneity, which deals with extant variation in the treatment effect across the population, causal moderation implies that counterfactual intervention upon a third (moderator) variable would alter the treatment effect. This study examines whether, when, and how researchers can actually investigate causal moderation rather than limit their
focus on treatment effect heterogeneity.

Evidence of causal moderation has important theoretical and policy implications that do not follow from the existence of treatment effect heterogeneity in general, and as such, researchers are often interested in making claims about causal moderation. Researchers focused on causal moderation, however, must confront additional estimation and identification hurdles when the moderator variable has not been randomized. This is a common situation in experimental social science, where financial, logistical, physical, psychological, or other constraints allow the researcher to randomize certain variables (the treatments) but not randomize other variables that are believed to interact with the treatment.

Yet, it is also common for researchers in this situation to make claims suggestive of causal moderation without fully grappling with the identification hurdles. Instead, treatment effect conditionality is most commonly analyzed on the basis of conventional subgroup analysis and/or underspecified interaction regression models. As this study will show, these standard approaches—namely, subsetting the treatment effect or interacting the treatment variable with the moderator of interest in a regression with control variables—do not constitute unbiased or consistent estimators of causal moderation effects given non-randomization of the moderator.

\[1\] For instance, one recent study on partisan identity and political emotion evaluates the extent to which political messages threatening a party’s electoral loss cause anger for copartisans (Huddy et al., 2015). In assessing whether the “partisan threat” experimental treatment leads to greater anger among respondents with stronger partisan identity, the authors describe this as the “effect of partisan identity...on emotional response to partisan threat and reassurance,” suggesting a causal moderation phenomenon. Another example can be found in a study on audience costs, which refers to the willingness of voters to punish elected officials for reneging on foreign policy commitments. The study finds varying degrees of audience costs—that is, an informational treatment describing a national policy decision as contradictory to international legal obligations is found to have varying effects on respondents’ disapproval of the decision—for respondents with different policy positions (Chaudoin, 2014). The author’s conclusion is that “audience costs are moderated by preferences over policy.” In another study evaluating the effects of legislative transparency measures on the performance and behavior of members of parliament in Vietnam, the authors assess variation in those effects across provinces with different levels of internet penetration (Malesky et al., 2012). The authors report that “each 10% increase in Internet penetration leads to an additional episode of speech by a treated delegate,” suggesting a causal dosage effect of internet penetration on the strength of the transparency treatment.
This study presents a generalized, non-parametric framework for estimating causal moderation effects given randomized treatments but non-randomized moderators, called the “parallel estimation framework,” that achieves a number of goals. First, it highlights how conventional approaches in the literature do not constitute unbiased or consistent estimators of causal moderation effects. Second, it offers researchers a simple, transparent approach for the estimation of causal moderation effects and lays out the assumptions under which this can be performed consistently and/or without bias. Third, as part of the estimation process, it allows researchers to implement their preferred method of covariate adjustment, including both parametric and non-parametric methods, or alternative identification strategies of their choosing. Fourth, it provides a set-up whereby sensitivity analysis designed for the average-treatment-effect context can be extended to the moderation context.

II. CAUSAL MODERATION IN CONTRAST TO TREATMENT EFFECT HETEROGENEITY

To illustrate what is at stake in properly estimating causal moderation, in contrast to other types of treatment effect conditionality, consider the following hypothetical experiment. Imagine a study where, in order to investigate a potential pathway for increasing local public goods provision in a developing country, researchers run a field experiment that randomly assigns certain towns to a program that provides specialized training to its civil servants. The researchers find that the treatment works—the training program improves public goods provision—but also that it appears to have differential effects for different towns. Specifically, the researchers suspect that the treatment is more effective in towns that have more civil servants per capita, i.e. towns with higher governmental capacity. Deciding how to extend their analysis and interpret their results then depends upon their research goals, and generally speaking, there may be two goals.
A. Treatment Effect Heterogeneity

As a first goal, the researchers may simply want to find treatment effect heterogeneity. In the example above, they may want to know the types of places where the training program is most effective, so that a future training policy can be rolled out in a targeted and cost-efficient manner. They do not necessarily need to know why it is more effective in certain places; they just need to be able to predict what those places are likely to be.

In this particular example, investigation of treatment effect heterogeneity across towns with different capacity levels could proceed using simple subgroup analysis. In addition, the researchers may want to identify other variables across which the treatment effect is heterogeneous but may not know which observed variables to focus attention on. Fortunately, recent years have seen an increase in interesting research into new methods to discover, estimate, and perform statistical inference on treatment effect heterogeneity beyond simple subgroup analysis, much of which integrates machine-learning techniques (Ding et al., 2016; Athey and Imbens, 2016; Wager and Athey, 2017; Beygelzimer and Langford, 2009; Dudík et al., 2011; Foster et al., 2011; Green and Kern, 2012; Imai and Ratkovic, 2013; Su et al., 2009; Zeileis et al., 2008; Weisberg and Pontes, 2015; Tian et al., 2014; Ratkovic and Tingley, 2017; Künzel et al., 2017; Powers et al., 2017).

B. Causal Moderation

In some cases, discovering the existence and extent of treatment effect heterogeneity may be the goal of a researcher’s inquiry into causal effect conditionality. However, the increasingly sophisticated methods for investigating treatment effect heterogeneity are not necessarily applicable to a second goal: estimating causal moderation effects. At stake here is the ability to claim that a moderator variable truly causes changes in the treatment effect.

The phenomenon of causal moderation is philosophically and formally anchored to the
notion of counterfactuals. Let \( Y \) denote an outcome variable of interest, \( T \) denote a binary treatment variable, and \( S \) denote a binary third variable that moderates the effect of \( T \) on \( Y \). Causal moderation refers to the effect that \( S \) causally exerts on \( T \)'s effect on \( Y \). To define this quantity of interest, consider a simple random sample of \( N \) subjects from a larger population, let \( Y_i(T_i, S_i) \) denote the potential outcomes for subject \( i = 1, 2, ..., N \), given values of the treatment \( T_i \) and moderator \( S_i \). The causal moderation estimand of interest will be referred to as the average treatment moderation effect (ATME) and denoted by \( \delta \):

**Definition 1:** The average treatment moderation effect (ATME) is defined as

\[
\delta = E\{(Y_i(1, S_i = 1) - Y_i(0, S_i = 1)) - (Y_i(1, S_i = 0) - Y_i(0, S_i = 0))\} = E\{(Y_i(1, 1) - Y_i(0, 1)) - (Y_i(1, 0) - Y_i(0, 0))\}
\]

Whereas treatment effect heterogeneity deals with variation in the treatment effect across different spaces of the existing distribution in the population, causal moderation implies that a counterfactual intervention upon a third (moderator) variable would alter the treatment effect. In the example above, the researchers may have found that the training program is more effective in high-capacity towns, but they did not randomly assign capacity and thus this result could be because capacity is related to any number of other things that are truly responsible for increasing the effectiveness of the treatment. If their goal is estimating causal moderation, then they need to know if it truly is higher capacity that causes the training program to be more effective.

The relationship between the realized \( Y_i \) and the potential outcomes for any subject \( i \) is \( Y_i = T_i S_i \cdot Y_i(1, 1) + T_i(1 - S_i) \cdot Y_i(1, 0) + (1 - T_i) S_i \cdot Y_i(0, 1) + (1 - T_i)(1 - S_i) \cdot Y_i(0, 0) \). Note that this study adopts a larger-population perspective, where the potential outcomes are fixed for any given subject \( i \) but the sample is a simple random sample of size \( N \) from a larger population, in contrast to the focus on average effects within a sample that is itself treated as the population of interest. See [Imbens and Rubin (2015)] for more details on this super-population approach and its statistical differences from (as well as practical similarities with) the finite sample-as-population perspective.
The answer to this question has theoretical and practical policy importance. If there is reason to believe that policy or other shocks in the future may lead to local capacity growth (or loss), then the researchers should be interested to know how this will likely impact the success of training programs that are introduced in the future. Alternatively, they may be interested to know if it would be cost-efficient to add funding for additional civil service employment to future training program roll-outs. To achieve this, finding treatment effect heterogeneity is not sufficient; what is needed is the estimation of causal moderation.

III. Identification and the Parallel Estimation Framework

This section introduces a generalized, non-parametric framework for estimating causal moderation effects given randomized treatments but non-randomized moderators, termed the “parallel estimation framework.” This framework provides a number of contributions to and addresses several problems in the methodological literature on interaction and moderation effects. By employing the non-parametric potential outcomes model to explicitly define and identify the estimand of interest, it highlights clearly and formally how conventional estimation approaches in the literature do not constitute consistent or unbiased estimators of causal moderation effects. More importantly, it offers researchers a simple, transparent approach that allows for the unbiased and/or consistent estimation of causal moderation effects.

A. Identification

As already presented above, the estimand of interest is the average treatment moderation effect (ATME):

$$
\delta = E\left[\{Y_i(1,1) - Y_i(0,1)\} - \{Y_i(1,0) - Y_i(0,0)\}\right]
$$

The problem, of course, is that for each subject, only one potential outcome is observable. Thus, a number of assumptions are necessary for identification of the ATME. The
assumptions will vary depending upon the precise identification strategy employed (this will be revisited later), but this section presents the most straightforward strategy in which selection on observables is applied to the moderator variable.

The first identification assumption is the stable unit treatment value assumption (SUTVA).

**Assumption 1 (SUTVA)**

If $T_i = T_i'$ and $S_i = S_i'$, then $Y_i(T, S) = Y_i(T', S')$, where $T$ and $S$ denote the full treatment and moderator vectors across subjects $i = 1, 2, ..., N$.

The second identification assumption is that the treatment vector is completely randomized.

**Assumption 2 (Completely Randomized Treatment)**

$P(T = a) = P(T = a')$ for all $a$ and $a'$ such that $i^T a = i^T a'$ where $i$ is the $N$-dimensional column vector with all elements equal to one.

The third identification assumption is that there are no individual-level effects of the treatment on the moderator $S$ or confounder variables $X$. Given the randomized treatment, this is guaranteed in the case that $S$ and $X$ are pre-treatment.

**Assumption 3 (Zero Effect of Treatment on Moderator and Confounders)**

$S_i(T_i = 1) = S_i(T_i = 0)$

$X_i(T_i = 1) = X_i(T_i = 0)$

for all $i = 1, 2, ..., N$.

The fourth identification assumption is conditional independence between the potential outcomes and the moderator $S$.

**Assumption 4 (Moderator Conditional Independence)**

$(Y_i(1, 1), Y_i(0, 1), Y_i(1, 0), Y_i(0, 0)) \perp S_i \mid X_i$
Note that given a randomized treatment and pre-treatment status of \( S \) and \( X \), the identification burden implied by Assumption 4 is no more stringent than the identification burden involved in the standard observational causal identification of an average treatment effect via selection on observables. To reflect this, the notation of Assumption 4 can be simplified as \( (Y_i(t, 1), Y_i(t, 0)) \perp \perp S_i \mid X_i \) for any \( t \).

In addition, identification via selection on observables requires the common support assumption, as is standard for observational causal identification.

**Assumption 5 (Moderator Common Support)**

\[ 0 < P(S_i = 1 \mid X_i) < 1 \]

The assumptions delineated above allow for identification of the ATME.

**Proposition 1:** Given Assumptions 1-5,

\[
\delta = E\left[ E[Y_i|T_i = 1, S_i = 1, X_i] - E[Y_i|T_i = 0, S_i = 1, X_i] \right.
\]

\[
- E[Y_i|T_i = 1, S_i = 0, X_i] + E[Y_i|T_i = 0, S_i = 0, X_i] \right] \]

**B. Moving from Estimand to Estimation**

The existence of an identification result, of course, does not necessarily ensure the existence of feasible or simple estimation methods. While clarifying the importance of the underlying assumptions, it is not immediately clear how to operationalize the identification result above for practical estimation of an ATME with empirical data. The following shows how the gap between identification and practical estimation can be bridged. More specifically, given the underlying assumptions, the estimand can be disaggregated into two separate components that can then be estimated separately, in parallel, and then recombined into the final ATME.

---

3See Appendix A in the Supplementary Materials (SM) for proofs of propositions.
estimate. This will be referred to as the parallel estimation framework, and its logic is illustrated by the following re-expression of the identification result:

\[ \delta = \]

\[ E \left[ Y_i | T_i = 1, S_i = 1, X_i \right] - E \left[ Y_i | T_i = 1, S_i = 0, X_i \right] \bigg| T_i = 1 \]

\[ - E \left[ Y_i | T_i = 0, S_i = 1, X_i \right] - E \left[ Y_i | T_i = 0, S_i = 0, X_i \right] \bigg| T_i = 0 \]

While presenting the estimand in this form is a simple mathematical re-expression, seeing it specifically in this form points to an opportunity for practical estimation of the estimand that is not immediately obvious. More specifically, whereas researchers would typically think to approach the task of estimating moderation effects by subsetting the units by moderator level, instead the units must be subsetted by treatment level, and covariate-adjustment estimation strategies can then be applied in parallel and separately for the treated units and control units to estimate \( E[Y_i(t, 1) - Y_i(t, 0)] \) for \( t = 0, 1 \). In other words,

\[ E \left[ Y_i | T_i = t, S_i = 1, X_i \right] - E \left[ Y_i | T_i = t, S_i = 0, X_i \right] \bigg| T_i = t \]

can be estimated separately for \( T_i = 0 \) and \( T_i = 1 \) by a covariate-adjustment estimator of the researcher’s choice. Notably, a variety of straightforward and widely used methods for performing this estimation already exist, such as regression and matching. In other words, this is equivalent to holding \( T \) constant and then estimating the effect of \( S \) on \( Y \) as one normally would in an observational study. Provided that the estimates of the effect of \( S \) on \( Y \) within each subset are unbiased and/or consistent, then their difference will constitute an unbiased and/or consistent estimate of the ATME, given the linearity of expectation and convergence in probability.

The reason this parallel estimation process works in practice is not immediately obvious just by looking at the identification result. A critical point that must be made is
that the ability to perform two separate estimation processes—the reason that the parallel estimation framework provides a viable bridge from estimand to estimation—is due to Assumptions 2 and 3. More specifically, this parallel estimation process works because the treatment has been randomly assigned and the moderator and control variables are unaffected by the treatment. The estimand of interest involves integration over the full distribution of $X$, which of course cannot be formally performed in the estimation process with empirical data. Instead, practical approximations of this integration process—such as regression, matching, etc.—must be performed with empirical data, and the viability of that approximation depends upon the distribution of the observed data. For the separation in the estimation process to not induce bias, it must be the case that the empirical distribution of $X$ arises from the same underlying population distribution for both subsets. That is, it must be that $f(X_i|T_i = 0) = f(X_i|T_i = 1) = f(X_i)$, where $f$ corresponds to the population. Because the treatment has been randomly assigned and the moderator and control variables are unaffected by the treatment, the joint distribution of $X$ and $S$ is the same for both treatment conditions, which means that estimation of $E\left[ E[Y_i|T_i = t, S_i = 1, X_i] - E[Y_i|T_i = t, S_i = 0, X_i] \bigg| T_i = t \right]$ can be performed separately for $t = 0, 1$. In contrast, subsetting the units by moderator level would lead to biased and inconsistent estimates of the ATME, as $f(X_i|S_i = 0) \neq f(X_i|S_i = 1)$ under the identification assumptions.

To appreciate the importance of the identification assumptions, consider the case in which the treatment is not randomly assigned but believed to adhere jointly to the same selection on observables as the moderator. In this case, the identification result would still hold, but the estimation process could not be broken into two separate parts in the manner suggested here because the joint distribution of the moderator and control variables would not be equivalent across subsets $T_i = 0, 1$. In other words, separating out the covariate adjustment
process into two separate tracts in this case would improperly adjust vis-à-vis the treatment, because the distributions upon which adjustment would be performed would be different for $T_i = 0, 1$. Hence, taking the difference between the two within-subset estimates would not recover an unbiased or consistent ATME estimate.

C. Alternative Identification Strategies

It is also important to note that there is nothing special about selection on observables as the within-subset identification strategy. Because of the randomization of the treatment and the symmetry of interaction effects, the data can be split into treatment-level subsets, and the causal effect of the moderator on the outcome within subset can be identified and estimated according to any identification strategy, provided that the relevant variables and additional model features are not affected by the treatment.

In other words, given Assumptions 1 and 2, identification of the ATME could be achieved via an alternative strategy if (a) the additional requisite assumptions are met such that the strategy achieves identification of the causal effect of $S$ on $Y$ in both subsets $T_i = 0$ and $T_i = 1$, and (b) the auxiliary variables and features used in the alternative identification strategy are not affected by the treatment. For instance, if a regression discontinuity (or encouragement/instrumental variables) design were used as the within-subset identification strategy, it would need to be the case that the forcing variable and threshold (or encouragement/instrument and principal strata) were pre-treatment or not otherwise impacted by the treatment. Of course, using such identification strategies would also require viewing the ATME as a local estimand and hence interpreting the corresponding estimate accordingly.

IV. Implementing the Parallel Estimation Framework

In sum, given a randomized treatment and non-randomized moderator, estimation of the ATME under the parallel estimation framework proceeds as follows:
1. Subset the data by treatment level.
   - *Not by moderator level,* which would result in biased and inconsistent estimation of the ATME.

2. Separately for each treatment subset $T_i = 0, 1$, estimate the effect of $S$ on $Y$.
   - Option A: Use preferred method of covariate adjustment (e.g. regression, matching, propensity score weighting, etc.) to perform estimation under the assumption that the effect of $S$ on $Y$ is identified conditional on observed covariates $X$.
   - Option B: Employ alternative identification strategy (e.g. regression discontinuity, instrumental variables, selection on unobservables) under alternative assumptions pertaining to the relationship between $S$ and $Y$.
   - In both cases, it must be true that $S$ and any auxiliary variables or features involved in identification are not affected by the treatment.

3. Calculate the difference between the two estimates from the previous step.
   - This is an estimate of the ATME ($\hat{\delta}$) by the symmetry of variable interactions. Provided that unbiased and/or consistent estimators of the effect of $S$ on $Y$ are employed in Step 2, and that $S$ and $X$ or any auxiliary variables/features involved in identification are not affected by the treatment, then $\hat{\delta}$ is also an unbiased and/or consistent estimate of the ATME by the linearity of expectation and convergence in probability.
   - Given simple random sampling from a larger population and the randomized treatment, $Var(\hat{\delta})$ can be approximated and estimated directly as the sum of the variances of the two estimates recovered in Step 2.

The following subsections will describe several concrete implementations of the parallel estimation framework and illustrate the problems with using standard approaches—subgroup analysis and underspecified interaction regressions—to estimate the ATME.

---

4In the theoretical case of simple random sampling from an infinite super-population, the two subsets $T_i = 0$ and $T_i = 1$ would be independent and hence $Var(\hat{\delta})$ would exactly equal the sum of the variances of the two estimates from Step 2.
A. Parallel Within-Treatment Regressions

The simplest method of estimation would be to use a set of parallel linear least squares regressions—one regression for the units with \( T_i = 0 \) and one regression for the units with \( T_i = 1 \). In each regression, the outcome variable \( Y \) is regressed on the moderator variable \( S \), along with the full set of covariates \( X \) conditional upon which it is assumed that \( S \) is independent of the potential outcomes. This then yields estimates of the average causal effect of \( S \) on \( Y \) within each treatment level. The difference between these two estimates then constitutes the estimate of the ATME, \( \hat{\delta}_{PR} \).

\[
Y_i = \hat{\alpha}_0 + \hat{\gamma}_0 S_i + X'_i \hat{\beta}_0 + \hat{\epsilon}_i \quad \forall \ i : T_i = 0 \\
Y_j = \hat{\alpha}_1 + \hat{\gamma}_1 S_j + X'_j \hat{\beta}_1 + \hat{\epsilon}_j \quad \forall \ j : T_j = 1 \\
\hat{\delta}_{PR} = \hat{\gamma}_1 - \hat{\gamma}_0
\]

The variance of the estimator \( \hat{\delta}_{PR} \) can be approximated as the sum of the variances of the linear least squares estimators \( \hat{\gamma}_0 \) and \( \hat{\gamma}_1 \):

\[
Var(\hat{\delta}_{PR}) \approx Var(\hat{\gamma}_0) + Var(\hat{\gamma}_1)
\]

B. Full Interaction Regression

The parallel regression equations also nest within a single regression equation. To see this, consider that the parallel regression equations can be combined as follows:

\[
Y_i = \hat{\alpha}_0 + (\hat{\alpha}_1 - \hat{\alpha}_0) T_i + \hat{\gamma}_0 S_i + X'_i \hat{\beta}_0 + (\hat{\gamma}_1 - \hat{\gamma}_0) T_i S_i + T_i X'_i (\hat{\beta}_1 - \hat{\beta}_0) + \hat{\epsilon}_i
\]

\[\text{This approach shares certain similarities with a framework proposed by Judd et al. (2001);} \text{ however, it is differently motivated and theoretically distinct. To begin, their context is one in which every individual is observed in the treatment and control condition, so the parallel regressions contain observations for all units in the sample. Furthermore, a more important difference is that the moderation implied by their parallel regression framework is descriptive, as they do not consider that the moderator may be correlated with any number of omitted variables that may actually be driving the difference in the treatment effect. In fact, the parallel regressions described there are simply an alternative, mathematically equivalent way of estimating a simple interaction between the treatment and moderator in the absence of controls, whereas the parallel regressions described here are presented specifically as a means of eliminating confounding of the causal moderation effect of interest given a linear functional form.}\]
This can be re-written in simpler terms as follows:

\[ Y_i = \hat{\alpha} + \hat{\tau}T_i + \hat{\omega}S_i + X_i'\hat{\beta} + \hat{\delta}_{FR}T_iS_i + T_iX_i'\hat{\xi} + \hat{\epsilon}_i \]

As a result, \( \hat{\delta}_{FR} = \hat{\delta}_{PR} \); the estimates are equivalent using both the parallel regression framework and the full interaction regression as specified above.

Note that \( T_iX_i' \) corresponds to the interactions between the treatment and all of the control variables. This highlights the erroneous nature of estimating an ATME in a single regression framework by merely controlling for the covariates by themselves and only including the interaction between the treatment and moderator, which is common among researchers seeking to estimate moderation effects. For instance, consider a data-generating process determined by the following model:

\[ Y_i = \alpha + \tau T_i + \omega S_i + \beta X_i + \delta T_iS_i + \xi T_iX_i + \epsilon_i \]

where \( \epsilon_i \) is distributed with mean zero and, for simplicity, \( X \) corresponds to a single covariate. This model corresponds to a situation where \( X \) not only affects the outcome itself but also, similar to the moderator of interest \( S \), interacts with the treatment’s effect on the outcome. Hence, it is not sufficient to simply include \( X \) as a regression control variable by itself in order to estimate the ATME as omission of the \( T \cdot X \) term would lead to omitted variable bias that affects \( \hat{\delta}_{FR} \). Given the omission of the \( T \cdot X \) term, the bias on \( \hat{\delta}_{FR} \) is a function of (a) the covariance between \( S \) and \( X \) and (b) the interaction effect between \( T \) and \( X \), denoted by \( \xi \) above. Depending upon the signs and magnitudes of \( \text{Cov}(S_i, X_i) \) and \( \xi \), this bias could lead to either amplification or attenuation of the ATME estimate, or even a sign reversal.

Another way to view the result above is that estimating an ATME in a single regression framework by merely controlling for the covariates alone would only be justified under the assumption that the full vectors of \( \beta_0 \) and \( \beta_1 \) from the parallel formulas are equal to each
other. This is tantamount to assuming that the relationship between the pre-treatment covariates and the outcome variable is not different across the treatment and control conditions. However, the enterprise of estimating causal moderation is motivated by the premise that one particular pre-treatment covariate of interest (the moderator) does have a different relationship with the outcome variable across the treatment and control conditions. To make the assumption that this is not true for the other pre-treatment covariates is not directly contradictory to the motivating premise, but it defies intuition and plausibility.

Instead, without the assumption that $\beta_0 = \beta_1$, the parallel regression estimation approach nests within a single regression if and only if the single regression includes the treatment, moderator, covariates, interaction between the treatment and moderator, and the full set of interactions between the treatment and the covariates.

C. Parallel Matching

The linear regression approaches described above rely upon specific functional form assumptions. This includes, of course, linearity. Less obvious, however, is that it also includes the assumption of constant effects given that linear regression performs a conditional-variance-weighting scheme that diverges from the type of weighting necessary to achieve an average effect estimate conditional on covariates (Angrist and Pischke 2009; Morgan and Winship 2015). These assumptions may be undesirable given certain data and when the estimation results are likely to be sensitive to functional form specifications. Thus, researchers may wish to avoid using linear regression as the covariate adjustment method in the parallel estimation process.

One alternative is to use parallel within-treatment-subset matching estimators to estimate the ATME. The parallel matching process is analogous to the parallel regression approach:

1. Split sample into treated and control sub-samples.
2. Within each sub-sample, match “moderated” units \( (S_i = 1) \) and “unmoderated” units \( (S_i = 0) \) on the confounding variables \( (X_i) \). Within each matched sub-sample, estimate the effect of \( S \) on \( Y \).

3. Compute the difference between these two estimated effects. This difference is an estimate of the \( ATME \).

As with the implementation of matching in the \( ATE \) context, standard diagnostics can be performed for both the treated and control sub-samples to assess common support and the quality of the matching process in producing covariate balance [Stuart, 2010].

D. Propensity Score Weighting

Another alternative approach can be undertaken via propensity score weighting. As with regression, this can be performed separately for each treatment-level subset, or the parallel paths can be combined into a single expression as follows:

**Proposition 2:** Given Assumptions [1-5],

\[
\hat{\delta} = E \left[ \frac{Y_i (T_i - p)(S_i - \pi(X_i))}{p(1-p)\pi(X_i)(1-\pi(X_i))} \right]
\]

where \( p \) is the probability of treatment and \( \pi(X_i) \) is the probability that \( S_i = 1 \) given covariate values \( X_i \).

This can be estimated by the following:

\[
\hat{\delta}_{PS} = \frac{1}{N} \sum_{i=1}^{N} Y_i (T_i - p)(S_i - \hat{\pi}(X_i)) \frac{p}{p(1-p)(1-\hat{\pi}(X_i))}
\]

where estimation of \( \hat{\pi}(X_i) \) can be performed using the researcher’s preferred propensity score estimation method.

E. Comparing the Parallel Estimation Framework to Conventional Estimation Approaches

The two standard approaches most commonly used in the literature to estimate moderation effects are subgroup analysis and underspecified interaction regressions. As the results
presented in this study demonstrate, these conventional methods generally do not constitute unbiased or consistent estimators of the ATME given a non-randomized moderator. This first approach, subgroup analysis, involves subsetting the data by different levels of the moderator variable, estimating the treatment effect within subsets, and calculating the difference between the two. This may be referred to alternatively as treatment-by-covariate interactions and comparison of conditional average treatment effects (CATEs) (cf. Gerber and Green, 2012). There are some warnings in the literature about the problems with interpreting effects estimated in this manner as causal moderation effects (except, of course, in the case where both the treatment and moderator are randomly assigned). Nonetheless, this simple technique is easy to understand and present, and empirical research performing this type of analysis will sometimes informally suggest a causal phenomenon.

The second common approach is using a controlled interaction regression, in which the outcome is regressed on the treatment, a third (moderator) variable of interest, the treatment-moderator interaction, and control variables. As has already been shown above, in spite of the addition of covariate adjustment, this method also falls short for estimating causal moderation in the case of a randomized treatment and non-randomized moderator. While this is not a new result in the literature (Yzerbyt et al., 2004; Hull et al., 1992), it appears not to be widely appreciated. Instead, empirical researchers continue to investigate causal moderation phenomena using interaction regressions that control for possible confounders but not their interactions with the treatment.

The controlled interaction regression specification can, of course, be extended for estimation of causal moderation by including additional interaction terms. That this appears to be done rarely suggests that, in spite of the large corpus of methodological literature on interaction terms in regression (e.g. Aiken and West, 1991; Brambor et al., 2006; Hainmueller et al., 2018; Kam and Franzese, 2009), there remains a lack of clarity on how to choose which
interaction terms to include in one’s regression model and which to omit. One problem is that the properties and practical usage of interaction/moderation effects are often discussed without explicitly stating what type of estimand (descriptive versus causal) is of interest. For more discussion on this, see VanderWeele (2015, pp. 268–270) and Kraemer et al. (2008). A second problem is that interaction/moderation effects are often discussed without reference to whether or not the treatment and moderator of interest are exogenous or randomized. The result is a lack of explicitness in terms of what types of confounding are possible and, accordingly, what types of adjustments must be made in response.

Furthermore, aside from researchers not properly dealing with confounding in their regression model specifications, an additional problem is the parametric structural nature of interaction regressions in general. Even in those cases where a causal moderation effect is explicitly posited, the parameterization of that effect within a specific statistical model such as a linear regression equation does not provide a general definition of the effect of interest and is associated with a number of modeling assumptions that may not be explicitly stated.

In sum, while there already exists extensive methodological literature on interaction and moderation effects, the disparate assumptions and analytical goals tacitly underlying different recommendations makes it difficult for researchers to know exactly what to do given their own specific data and goals. Furthermore, because most of the literature is anchored within...
the regression context, the recommendations often lack general formal results to justify certain specifications; researchers must instead fall back on the usual, slightly uncomfortable assumption that the posited structural model is correct. This is not reassuring for researchers who want a robust model that is both motivated by clear-cut theoretical guidance yet also acknowledges that the estimation models we specify are rarely correct or complete. By formalizing the definition of a causal moderation effect in the form of the ATME, this study seeks to clarify precisely what is meant by causal moderation, what assumptions its identification requires, and how its estimation can be undertaken given those assumptions.

V. Application: Asylum Seeker Conjoint Experiment

European governments have recently struggled with the largest refugee crisis since World War II, with Europe receiving approximately 1.3 million new asylum claims in 2015. Exacerbating the political difficulties of accommodating asylum seekers while minimizing domestic public backlash is that, while most asylum seekers currently originate from Muslim-majority countries, research has shown that European voters would prefer to grant asylum to non-Muslim applicants (Bansak et al., 2016). Furthermore, newly empowered right-wing parties have sought to mobilize citizens around asylum policy issues, appealing to both nationalistic impulses and Islamophobia in promoting anti-asylum policies. It is possible that these parties and other political actors have exacerbated the anti-Muslim bias toward asylum seekers by stoking nationalism across the continent.

This episode is important in and of itself, and it also represents an example of a long-standing theme explored by political scientists whereby nationalism is considered a malleable dimension of public opinion that may fluctuate over time, is endogenous to current events, and constitutes a potential target for manipulation via government intervention and elite political communication (e.g. Mueller, 1970; Baker and Oneal, 2001; Weiss, 2013). In this
particular case, we may ask: to what extent might nationalism increase Islamophobia? More specifically, does increased nationalism cause a stronger preference for non-Muslim asylum seekers?

This section presents an application of the parallel estimation framework using data from a recent conjoint experiment embedded in an online survey fielded in fifteen European countries with approximately 18,000 respondents (Bansak et al., 2016). In the conjoint experiment, respondents were presented with pairs of hypothetical asylum-seeker profiles comprised of various attributes with randomly assigned levels. The analysis here focuses on one attribute in particular, which serves as the randomly assigned treatment $T$ in the parallel estimation process: whether the asylum seeker was Muslim (1) or Christian (0). For each pair of asylum seekers, the respondent was asked to specify which s/he preferred, providing a binary preference outcome variable $Y$. Each respondent’s level of nationalism was also measured via a set of questions in the survey combined into an index. This measure is dichotomized into a binary measure of high-versus-low nationalism, with the median index value chosen as the cut point, for use as the moderator variable $S$. Thus, being investigated is whether nationalism moderates the effect of an asylum seeker’s religion on voters’ acceptance of the asylum seeker. Finally, a range of demographic and political ideological characteristics were measured, which are used as the set of control variables $X$. More details on the study design and variables can be found in Appendix B in the SM.

Figures 1 and 2 present the results, pooled and across countries, with moderation effect point estimates and 95% confidence intervals estimated using the conventional approaches and parallel estimation framework respectively. The results from the conventional estimators, shown in Figure 1, suggest relatively robust moderation. The first estimator is the difference in subset effects. The two subsets are high-nationalism respondents and low-nationalism

---

This includes: gender, age, income, education, and left-right political ideology.
respondents, and each subset effect is the simple difference between the probability of accepting an asylum seeker in the treatment condition (i.e. given a Muslim asylum seeker) and the probability of accepting an asylum seeker in the control condition (i.e. given a Christian asylum seeker). The differences in subset effects highlight heterogeneity of voters’ preferences regarding the religion of asylum seekers: the preference for Christian asylum seekers over Muslim asylum seekers is stronger among high-nationalism respondents than low-nationalism respondents in all countries, a statistically significant finding at $\alpha = 0.05$ in ten out of the fifteen countries included in the survey.\(^8\) Substantively, the magnitude of that heterogeneity varies by country. For instance, in Austria, the difference in subset effects is about $-0.1$, meaning that the negative gap in the probability of accepting a Muslim asylum seeker relative to a Christian asylum seeker grows by 10 additional percentage points when moving from the low-nationalism segment of Austrian voters to the high-nationalism segment.

\(^8\)Standard errors are clustered by respondent, since each respondent evaluated multiple pairs of profiles.
The second estimator whose results are presented in Figure 1 is the controlled interaction regression estimator, which is the estimate of the interaction effect between the treatment and moderator in a regression of the outcome on the treatment, moderator, treatment-moderator interaction, and control variables. While a number of the control variables are correlated with nationalism (as should be expected), and while we should also expect a number of these control variables to bear upon voters’ biases toward asylum seekers, the figure shows that the inclusion of controls in the interaction regression has virtually no effect on the moderation effect estimates, compared to the simple difference in subset effects. This may seem surprising, unless one takes into account the identification result and corresponding estimation strategies presented in this study, and considers that the interaction regression does not include interactions between the treatment and controls.

As explained in this study, it would be incorrect to interpret the estimates in Figure 1 as causal moderation effects because they are based on estimators that do not perform covariate adjustment in a way that allows for estimation of a clearly defined and identified causal moderation estimand. In contrast, results using the parallel regression method and the parallel matching method are reported in Figure 2. As can be seen, these results paint a much more conservative picture in terms of the extent to which increased nationalism can strengthen Islamophobic biases toward asylum seekers. Specifically, with only a few exceptions the causal moderation effect estimates yielded by these methods are attenuated toward zero compared to the conventional estimates. Of the ten countries where the conventional estimators yielded statistically significant moderation effects, only one is found to have statistically significant causal moderation effects according to both the parallel regression and parallel matching estimators, and only four others are found to have statistically significant

9 The matching procedure used Mahalanobis distance and matched all of the moderated (high nationalism) units with the closest unmoderated units, with replacement.
causal moderation effects by either estimator. Given this systematically consistent attenuation across countries, these results indicate that a substantial portion of the moderation effect attributed to nationalism by the conventional estimators is actually the result of nationalism being correlated with other variables that themselves impact the Islamophobic bias, namely left-right political ideology, education, and age.

In sum, this application highlights how estimates of ostensible moderation effects can vary substantially depending on whether the analysis follows conventional practices or implements parallel framework estimators built specifically to estimate a well-defined and identified causal moderation effect (i.e. the ATME). In this example, methods suited for uncovering heterogeneity of effects convincingly found the preference for non-Muslim asylum seekers to vary across voters with different levels of nationalism. This uncovering of key subpopulations may be interesting and important, but it does not provide more policy and action-oriented evidence on the extent to which varying (and the possibility of political actors strategically manipulating) levels of nationalism would cause an increase or decrease in the preference for non-Muslim asylum seekers. To accomplish the latter task, the ATME must be estimated. Appendix C in the SM presents additional applications illustrating the usage of the parallel estimation framework and comparing it to conventional estimation approaches.

VI. Discussion and Conclusions

A. Sensitivity Analysis

Identification of the ATME via selection on observables relies on the assumption that all variables that confound the relationship between the moderator and the outcome are conditioned upon. Of course, as with the estimation of simple treatment effects using observational data, it is impossible to know and difficult to argue that every possible confounder has been observed. Instead, the practical question is: How serious is the bias likely to be due to unobserved confounders, and how sensitive is the estimate of interest—the ATE in
the simple context, and the ATME in the context of this study—to the existence of such remaining bias? Answering this question can be done by pairing subject-matter expertise with a rigorous sensitivity analysis.

In addition to being easy and transparent to implement, the parallel estimation framework also allows for an adaptation of ATE sensitivity analysis to the causal moderation context. Appendix D in the SM develops and presents a sensitivity analysis for estimation of the ATME using the parallel estimation framework, along with an application.

B. Scope Conditions for Causal Moderation

Causal moderation, as described and formally defined in this study, is a causal phenomenon in which the counterfactual intervention upon a moderator variable would alter the effect of the treatment on the outcome. Accordingly, for the ATME as conceptualized in this study to be a relevant and fundamentally meaningful estimand of interest, it must be the case not only that the moderator in question is unaffected by the treatment but also that the moderator is indeed mutable at the individual level. If a moderator is not actually mutable, the ATME does not represent a conceptually clear causal quantity, just as a standard ATE does not have clear meaning for a treatment variable that is not mutable.

Thus, for researchers interested in investigating causal moderation and estimating an ATME for a particular moderator, they must first ask themselves whether such a goal makes sense given the nature of the moderator variable. Moderators that are immutable characteristics, such as the research subjects’ racial identities, may fall out of the scope of causal moderation and should instead be investigated within the estimation and interpretive frameworks of treatment effect heterogeneity.\textsuperscript{10} In contrast, investigation and estimation of ATMEs, for either theoretical or policy objectives, is a fruitful endeavor for moderators that

\textsuperscript{10}Note, however, that in some studies it is not racial identity but rather perception of racial identity that is of interest, and the latter can be manipulated.
are reasonably changeable at the individual level, as such moderators could be manipulated or intervened upon by external forces.

In addition, to guard against data dredging and erroneous statistical inferences, moderators of interest should ideally be identified \textit{a priori} before the researcher has begun analysis, and multiple testing corrections should be made where appropriate.

\subsection*{C. Conclusions}

If a researcher’s goal is to estimate causal moderation effects, it is generally preferable to design a factorial experiment in which the treatment and moderator are both randomized. Unfortunately, however, randomization of the moderator is often not possible, in which case identification of causal moderation effects requires additional work in the analysis stage. For researchers who find themselves in such a situation, this study has introduced a generalized, non-parametric, and straightforward framework for estimating causal moderation effects.

The framework presented in this study makes a number of contributions. First, by employing the potential outcomes model to explicitly define and identify the estimand of interest, it highlights clearly and formally how conventional estimation approaches in the literature do not constitute consistent or unbiased estimators of causal moderation effects. Second, it offers researchers a simple, transparent approach for the estimation of causal moderation effects and lays out the assumptions under which this can be performed consistently and/or without bias. Third, as part of the estimation process, it allows researchers to implement their preferred method of covariate adjustment, including both parametric and non-parametric methods, or alternative identification strategies of their choosing. Fourth, it provides a set-up whereby sensitivity analysis designed for the ATE context can be extended to the causal moderation context.

Ultimately, the value of investigating causal moderation is in allowing researchers to
better understand and anticipate the effectiveness of a particular treatment should other variables (i.e. moderators) change or be intervened upon when the treatment is administered or investigated in other contexts and/or in the future. This endeavor is core to the broader, critical, but often elusive goal of understanding the generalizability of treatment effects.

References

Abramowitz, A. I. (1994). Issue evolution reconsidered: Racial attitudes and partisanship in the U.S. electorate. American Journal of Political Science, 38(1):1–24.

Aiken, L. S. and West, S. G. (1991). Multiple Regression: Testing and Interpreting Interactions. Thousand Oaks, CA: SAGE Publications, Inc.

Angrist, J. D. and Pischke, J.-S. (2009). Mostly Harmless Econometrics: An Empiricist’s Companion. Princeton: Princeton University Press.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27):7353–7360.

Baker, W. D. and Oneal, J. R. (2001). Patriotism or opinion leadership? The nature and origins of the “rally ‘round the flag” effect. Journal of Conflict Resolution, 45(5):661–687.

Bansak, K., Hainmueller, J., and Hangartner, D. (2016). How economic, humanitarian, and religious concerns shape European attitudes toward asylum seekers. Science, 354(6309):217–222.

Beygelzimer, A. and Langford, J. (2009). The offset tree for learning with partial labels. Technical report, ACM SIGKDD. https://arxiv.org/abs/0812.4044.

Brambor, T., Clark, W. R., and Golder, M. (2006). Understanding interaction models: Improving empirical analyses. Political Analysis, 14:63–82.

Carmines, E. G. and Stimson, J. A. (1989). Issue Evolution: Race and the Transformation of American Politics. Princeton: Princeton University Press.

Chaplin, W. F. (1991). The next generation in moderation research in personality psychology. Journal of Personality, 59:143–178.

Chaudoin, S. (2014). Promises or policies? An experimental analysis of international agreements and audience reactions. International Organization, 68(01):235–256.

Cohen, J. and Cohen, P. (1983). Applied Multiple Regression/Correlation Analysis for the
Ding, P., Feller, A., and Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):655–671.

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. Technical report, International Machine Learning Society (IMLS). [https://arxiv.org/abs/1103.4601](https://arxiv.org/abs/1103.4601).

Federico, C. M. (2004). When do welfare attitudes become racialized? The paradoxical effects of education. *American Journal of Political Science*, 48(2):374–391.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880.

Frazier, P. A., Tix, A. P., and Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology*, 51(1):115–134.

Gerber, A. S. and Green, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton & Company.

Green, D. P. and Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public Opinion Quarterly*, 76(3):491–511.

Hainmueller, J., Mummolo, J., and Xu, Y. (2018). How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice. *Political Analysis (conditionally accepted)*. [https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2739221](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2739221).

Henry, P., Reyna, C., and Weiner, B. (2004). Hate welfare but help the poor: How the attributional content of stereotypes explains the paradox of reactions to the destitute in America. *Journal of Applied Social Psychology*, 34(1):34–58.

Huddy, L., Mason, L., and Aarøe, L. (2015). Expressive partisanship: Campaign involvement, political emotion, and partisan identity. *American Political Science Review*, 109(01):1–17.

Hull, J. G., Tedlie, J. C., and Lehn, D. A. (1992). Moderator variables in personality research: The problem of controlling for plausible alternatives. *Personality and Social Psychology Bulletin*, 18:115–117.

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7:443–470.
Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

ISSP Research Group (2012). International Social Survey Programme: National Identity II- ISSP 2003. *GESIS Data Archive, Cologne*, 2. [https://www.gesis.org/issp/modules/issp-modules-by-topic/national-identity/2003/](https://www.gesis.org/issp/modules/issp-modules-by-topic/national-identity/2003/)

Judd, C. W., Kenny, D. A., and McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods*, 6(2):115–134.

Kam, C. D. and Franzese, Jr., R. J. (2009). *Modeling and Interpreting Interactive Hypotheses in Regression Analysis*. Ann Arbor: University of Michigan Press.

Kluegel, J. R. and Smith, E. R. (1986). *Beliefs About Inequality: Americans’ Views of What Is and What Ought to Be*. New York: Aldine de Gruyter.

Kraemer, H., Kiernan, M., Essex, M., and Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology*, 27:101–108.

Künzel, S., Sekhon, J., Bickel, P., and Yu, B. (2017). Meta-learners for estimating heterogeneous treatment effects using machine learning. Technical report. [https://arxiv.org/abs/1706.03461](https://arxiv.org/abs/1706.03461)

Malesky, E., Schuler, P., and Tran, A. (2012). The adverse effects of sunshine: A field experiment on legislative transparency in an authoritarian assembly. *American Political Science Review*, 106(04):762–786.

Mansfield, E. D. and Mutz, D. C. (2009). Support for free trade: Self-interest, sociotropic politics, and out-group anxiety. *International Organization*, 63(3):425–457.

McClelland, G. H. and Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114:376–390.

Morgan, S. L. and Winship, C. (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Second Edition)*. New York: Cambridge University Press.

Mueller, J. E. (1970). Presidential popularity from Truman to Johnson. *American Political Science Review*, 64(1):18–34.

Olken, B. A. (2007). Monitoring corruption: Evidence from a field experiment in Indonesia. *Journal of Political Economy*, 115(2):200–249.
Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., and Tibshirani, R. (2017). Some methods for heterogeneous treatment effect estimation in high-dimensions. Technical report. https://arxiv.org/abs/1707.00102

Rasinski, K. A. (1989). The effect of question wording on public support for government spending. Public Opinion Quarterly, 53(3):388–394.

Ratkovic, M. and Tingley, D. (2017). Sparse estimation and uncertainty with application to subgroup analysis. Political Analysis, 25(1):1–40.

Shaw, G. M. and Shapiro, R. Y. (2002). Trends: Poverty and public assistance. Public Opinion Quarterly, 66(1):105–128.

Smith, T. W. (1987). That which we call welfare by any other name would smell sweeter. Public Opinion Quarterly, 51(1):75–83.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical Science, 25(1):1–21.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup analysis via recursive partitioning. The Journal of Machine Learning Research, 10:141–158.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. Journal of the American Statistical Association, 109(508):1517–1532.

VanderWeele, T. J. (2015). Explanation in Causal Inference: Methods for Mediation and Interaction. New York: Oxford University Press.

Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association (forthcoming). https://doi.org/10.1080/01621459.2017.1319839

Weisberg, H. I. and Pontes, V. P. (2015). Post hoc subgroups in clinical trials: Anathema or analytics? Clinical Trials, 12(4):357–364.

Weiss, J. C. (2013). Authoritarian signaling, mass audiences, and nationalist protest in China. International Organization, 67(1):1–35.

Yzerbyt, V. Y., Muller, D., and Judd, C. M. (2004). Adjusting researchers’ approach to adjustment: On the use of covariates when testing interactions. Journal of Experimental Social Psychology, 40:424–431.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. Journal of Computational and Graphical Statistics, 17(2):492–514.
Supplementary Materials

for

A Generalized Framework for the Estimation of Causal Moderation Effects with Randomized Treatments and Non-Randomized Moderators

Kirk Bansak
APPENDIX A: PROOFS

Proof of Proposition 1

To begin, note that Assumptions 2, 3, and 5 together ensure joint common support:

\[ 0 < P(T_i = t, S_i = s|X_i) < 1 \]

for any \( s = 0,1 \) and \( t = 0,1 \). This follows from Assumption 5 that \( 0 < P(S_i = s|X_i) < 1 \), and Assumptions 2 and 3 implying that \( T_i \perp (S_i, X_i) \) and hence that \( 0 < P(T_i = t|X_i) < 1 \) and \( P(T_i = t, S_i = s|X_i) = P(T_i = t|X_i)P(S_i = s|X_i) \). Thus, \( 0 < P(T_i = t, S_i = s|X_i) < 1 \).

Joint common support then allows for defining \( E[Y_i|T_i = t, S_i = s, X_i] \) for all \( t = 0,1 \) and \( s = 0,1 \), which thereby allows for defining:

\[
\begin{align*}
\{E[Y_i|T_i = 1, S_i = 1, X_i] - E[Y_i|T_i = 0, S_i = 1, X_i]\} \\
\{E[Y_i|T_i = 1, S_i = 0, X_i] - E[Y_i|T_i = 0, S_i = 0, X_i]\} = E[Y_i|T_i = 1, S_i = 1, X_i] - E[Y_i|T_i = 0, S_i = 1, X_i] \\
- E[Y_i|T_i = 1, S_i = 0, X_i] + E[Y_i|T_i = 0, S_i = 0, X_i]
\end{align*}
\]

Given \( Y_i = T_iS_i \cdot Y_i(1,1) + T_i(1 - S_i) \cdot Y_i(1,0) + (1 - T_i)S_i \cdot Y_i(0,1) + (1 - T_i)(1 - S_i) \cdot Y_i(0,0) \), this equals:

\[
\begin{align*}
E[Y_i(1,1)|T_i = 1, S_i = 1, X_i] - E[Y_i(0,1)|T_i = 0, S_i = 1, X_i] \\
- E[Y_i(1,0)|T_i = 1, S_i = 0, X_i] + E[Y_i(0,0)|T_i = 0, S_i = 0, X_i]
\end{align*}
\]

Now, note that Assumptions 2 and 3 together imply that

\[
(Y_i(1,1), Y_i(0,1), Y_i(1,0), Y_i(0,0), S_i, X_i) \perp T_i
\]

which means that for any \( t = 0,1 \) and \( s = 0,1 \), the conditional probability distribution of \( Y_i(t, s) \) has the following properties:

\[
f(Y_i(t,s)|T_i = t, S_i = s, X_i) = \frac{f(Y_i(t,s), T_i = t, S_i = s, X_i)}{f(T_i = t, S_i = s, X_i)} = \frac{f(Y_i(t,s), S_i = s, X_i)P(T_i = t)}{f(S_i = s, X_i)P(T_i = t)}
\]

\[
= \frac{f(Y_i(t,s), S_i = s, X_i)}{f(S_i = s, X_i)} = f(Y_i(t,s)|S_i = s, X_i)
\]

and hence

\[
E[Y_i(t,s)|T_i = t, S_i = s, X_i] = E[Y_i(t,s)|S_i = s, X_i]
\]
Further, note that for any $s = 0, 1$, given Assumption 4,

$$E[Y_i(t, s)|S_i = s, X_i] = E[Y_i(t, s)|X_i]$$

All of this implies that

$$E[Y_i(1, 1)|T_i = 1, S_i = 1, X_i] - E[Y_i(0, 1)|T_i = 0, S_i = 1, X_i]$$

$$-E[Y_i(1, 0)|T_i = 1, S_i = 0, X_i] + E[Y_i(0, 0)|T_i = 0, S_i = 0, X_i]$$

$$= E[Y_i(1, 1)|X_i] - E[Y_i(0, 1)|X_i] - E[Y_i(1, 0)|X_i] + E[Y_i(0, 0)|X_i]$$

Hence,

$$\delta =$$

$$E[Y_i(1, 1)] - E[Y_i(0, 1)] - E[Y_i(1, 0)] + E[Y_i(0, 0)]$$

$$= E \left[ E[Y_i(1, 1)|X_i] - E[Y_i(0, 1)|X_i] - E[Y_i(1, 0)|X_i] + E[Y_i(0, 0)|X_i] \right]$$

$$= E \left[ E[Y_i|T_i = 1, S_i = 1, X_i] - E[Y_i|T_i = 0, S_i = 1, X_i]$$

$$-E[Y_i|T_i = 1, S_i = 0, X_i] + E[Y_i|T_i = 0, S_i = 0, X_i] \right]$$

$\square$
Proof of Proposition 2

\[ E[Y_i|T_i = 1, S_i = 1, X_i] - E[Y_i|T_i = 1, S_i = 0, X_i] + E[Y_i|T_i = 0, S_i = 0, X_i] - E[Y_i|T_i = 0, S_i = 1, X_i] \]

\[ = E \left[ \frac{Y_i}{\pi_{11}(X_i)} \right] | T_i = 1, S_i = 1, X_i \]  
\[ \pi_{11}(X_i) + E \left[ -\frac{Y_i}{\pi_{10}(X_i)} \right] | T_i = 1, S_i = 0, X_i \]  
\[ \pi_{10}(X_i) \]

+ \[ E \left[ \frac{Y_i}{\pi_{00}(X_i)} \right] | T_i = 0, S_i = 0, X_i \]  
\[ \pi_{00}(X_i) + E \left[ -\frac{Y_i}{\pi_{01}(X_i)} \right] | T_i = 0, S_i = 1, X_i \]  
\[ \pi_{01}(X_i) \]

where \( \pi_{ts}(X_i) \) refers to the probability of a unit receiving \( T_i = t \) and \( S_i = s \) as a function of \( X_i \).

Given Assumptions 2 and 3:

\[ P(T_i = t, S_i = s|X_i) = P(T_i = t)P(S_i = s|X_i) \]

Let \( P(T_i = 1) = p \), \( P(T_i = 0) = 1 - p \), \( P(S_i = 1|X_i) = \pi(X_i) \), \( P(S_i = 0|X_i) = 1 - \pi(X_i) \),

which simplifies the expression above to:

\[ E \left[ \frac{Y_i}{p\pi(X_i)} \right] | T_i = 1, S_i = 1, X_i \]  
\[ p\pi(X_i) \]

+ \[ E \left[ \frac{-Y_i}{p(1 - \pi(X_i))} \right] | T_i = 1, S_i = 0, X_i \]  
\[ p(1 - \pi(X_i)) \]

+ \[ E \left[ \frac{Y_i}{(1 - p)(1 - \pi(X_i))} \right] | T_i = 0, S_i = 0, X_i \]  
\[ (1 - p)(1 - \pi(X_i)) \]

+ \[ E \left[ \frac{-Y_i}{(1 - p)\pi(X_i)} \right] | T_i = 0, S_i = 1, X_i \]  
\[ (1 - p)\pi(X_i) \]

which, using the conditional values and the law of total expectation, becomes:

\[ E \left[ Y_i \frac{(T_i - p)(S_i - \pi(X_i))}{p(1 - p)\pi(X_i)(1 - \pi(X_i))} \right] \]

Taking the expectation, we achieve:

\[ E\left[ E[Y_i|T_i = 1, S_i = 1, X_i] - E[Y_i|T_i = 1, S_i = 0, X_i] \right. \]

\[ + E[Y_i|T_i = 0, S_i = 0, X_i] - E[Y_i|T_i = 0, S_i = 1, X_i] \]

\[ = \delta \]
\[E[Y_i|T_i = 1, S_i = 1] - E[Y_i|T_i = 1, S_i = 0] + E[Y_i|T_i = 0, S_i = 0] - E[Y_i|T_i = 0, S_i = 1]\]
\[= E \left[ E \left[ Y_i \frac{(T_i - p)(S_i - \pi(X_i))}{p(1 - p)(\pi(X_i)(1 - \pi(X_i)))} | X_i \right] \right] \]
\[= E \left[ Y_i \frac{(T_i - p)(S_i - \pi(X_i))}{p(1 - p)(\pi(X_i)(1 - \pi(X_i)))} \right] \]

□
Appendix B: Asylum Seeker Application Details

This appendix provides additional details on the application presented in the main text, which is an extension of a previous study (Bansak et al., 2016).

Figure B1 displays an example of a pair of profiles shown to respondents, and Table B1 describes the conjoint attributes. The following is the text used to introduce the respondents to the conjoint section of the survey, as well as the questions used to measure the respondents’ evaluations of the asylum-seeker profiles and hence construct the outcome variable.

Prelude:

“Now we would like to show you the profiles of potential applicants for asylum in Europe. You will be shown pairs of asylum seekers, along with several of their attributes. We would like to know your opinion regarding whether you would be in favor of sending each applicant back to their country of origin or allowing them to stay in [Respondent’s Country].

In total, we will show you five comparison pairs. Please take your time when reading the descriptions of each applicant. People have different opinions about this issue, and there are no right or wrong answers.”

Questions:

1. (Rating) “On a scale from 1 to 7, where 1 indicates that [Respondent’s Country] should absolutely send the applicant back to their country of origin and 7 indicates that [Respondent’s Country] should definitely allow the applicant to stay, how would you rate each of the asylum seekers described above?”

2. (Choice) “Now imagine that you had to choose one applicant who would be allowed to stay in [Respondent’s Country], and the other applicant would be sent back to their own country of origin. Which of the two applicants would you personally prefer to be allowed to stay in [Respondent’s Country]?”

The Choice outcome variable is used for the analysis presented in the main text. In addition to the conjoint component, the survey instrument also contained a number of questions that measured various characteristics and attitudes of the respondents. Tables B2-B6 describe these additional variables and present summary statistics.
Figure B1: Example profile pair

This figure displays an example of a pair of asylum-seeker profiles seen by the survey respondents. This example comes from the English version of the survey administered to respondents in the United Kingdom. Each respondent saw and evaluated five separate pairs. The order of the attributes (rows) was fully randomized between respondents, but for each respondent, the order was kept constant across the five pairs they were shown. The specific attribute levels (values in the cells in the last two columns) were fully randomized between and within respondents. Source: Bansak et al. (2016).
Table B1: Conjoint attributes and attribute values/levels

This table displays the attributes and attribute levels used to construct the asylum-seeker profiles.

| Attribute                  | Attribute Levels                                                                 |
|----------------------------|----------------------------------------------------------------------------------|
| Asylum Testimony           | No inconsistencies, Minor inconsistencies, Major inconsistencies                  |
| Gender                     | Female, Male                                                                     |
| Country of Origin          | Syria, Afghanistan, Kosovo, Eritrea, Pakistan, Ukraine, Iraq                      |
| Age                        | 21 years, 38 years, 62 years                                                     |
| Previous Occupation        | Unemployed, Cleaner, Farmer, Accountant, Teacher, Doctor                         |
| Vulnerability              | None, Post-traumatic stress disorder (PTSD), Victim of torture, No surviving family members, Physically handicapped |
| Reason for Migrating       | Persecution for political views, Persecution for religious beliefs, Persecution for ethnicity, Seeking better economic opportunities |
| Religion                   | Christian, Agnostic, Muslim                                                       |
| Language Skills            | Speaks fluent [language of respondent’s country], Speaks broken [language of respondent’s country], Speaks no [language of respondent’s country] |
This table describes the outcome variable and other respondent characteristics measured in the survey.

| Variable       | Description                                                                                                                                 |
|----------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| **Forced Choice** | This is a binary indicator for whether or not a profile was the preferred profile in its respective pair. It is based on respondents’ answers to the following question: “Now imagine that you had to choose one applicant who would be allowed to stay in [Respondent’s Country], and the other applicant would be sent back to their own country of origin. Which of the two applicants would you personally prefer to be allowed to stay in [Respondent’s Country]?” |
| **Age**         | Self-reported age. For the analysis used in the main text, indicators for four age bins were employed: 18-29, 30-44, 45-64, 65+.                                                                          |
| **Gender**      | Self-reported gender.                                                                                                                         |
| **Education**   | Self-reported highest level of education achieved. The customized country-specific educational level questionnaire options used by the European Social Survey were employed for each country, and the respondents’ educational levels were mapped onto the harmonized European version of the International Standard Classification of Education (EISCED) scale. The scale contains seven major categories: less than lower secondary (1), lower secondary (2), lower tier upper secondary (3), upper tier upper secondary (4), advanced vocational/sub-degree (5), lower tertiary education/BA level (6), and higher tertiary education/≥ MA level (7). |
| **Income**      | Self-reported household income. Respondents’ household income levels were measured using the customized country-specific household income decile bins constructed in the European Social Survey (the most recent available version of the European Social Survey was used for each country in the sample). For the analysis in the main text, these deciles were coarsened into quintiles. |
Table B3: Variables, continued

| Variable         | Description                                                                                                                                 |
|------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| Political ideology | Self-reported placement on the left-right political ideological spectrum. Ideological placement was measured using a standard self-identification scale ranging from 0 on the left to 10 on the right. The question wording used to measure the respondents’ ideology was as follows: “In politics people often talk of ‘left’ and ‘right’. On this scale from 0 (left) to 10 (right), where would you classify your own political views?” For the analysis in the main text, political ideology was coarsened into five bins: far left (0-2), left (3-4), center (5), right (6-7), and far right (8-10). |
| Nationalism      | Multi-construct index of nationalism. To measure nationalism, the survey employed four commonly used prompts—taken from the National Identity module of the International Social Survey Programme ([ISSP Research Group](https://www.issp.org) [2012]) and [Mansfield and Mutz (2009)](https://doi.org/10.1016/j.jpubeco.2009.07.003)—tapping into different dimensions of nationalism. For each prompt, respondents specified their level of (dis)agreement: agree strongly (2), agree (1), neither agree nor disagree (0), disagree (-1), disagree strongly (-2). These values were then summed to create an index for nationalism, and the median value was used as a cut point for creating a dichotomized version of the variable (high-versus-low nationalism) for the analysis in the main text. The prompts included the following. (a) “I would rather be a citizen of [Respondent’s Country] than of any other country in the world.” (b) “Generally speaking [Respondent’s Country] is a better country than most other countries.” (c) “[Respondent’s Country] should follow its own interests, even if this leads to conflicts with other nations.” (d) “[Respondent’s Country’s] government should just try to take care of the well-being of [Respondent’s Country’s] citizens and not get involved with other nations.” |
Table B4: Number of respondents per country

| Country               | Sample Size |
|-----------------------|-------------|
| Austria               | 1206        |
| Czech Republic        | 1202        |
| Denmark               | 1201        |
| France                | 1203        |
| Germany               | 1200        |
| Greece                | 1200        |
| Hungary               | 1200        |
| Italy                 | 1200        |
| Netherlands           | 1200        |
| Norway                | 1202        |
| Poland                | 1201        |
| Spain                 | 1203        |
| Sweden                | 1203        |
| Switzerland           | 1208        |
| United Kingdom        | 1201        |
| **Total**             | **18030**   |
Table B5: Summary statistics

This table displays the distribution of respondents along key demographic and ideological dimensions, by country.

| Gender | Age | Income Quintile | Political Ideology |
|--------|-----|-----------------|--------------------|
|        | Female | 29/under | 30-39 | 40-49 | 50-59 | 60/over | First | Second | Third | Fourth | Fifth | Left | Center | Right |
|--------|--------|-----------|-------|-------|-------|--------|-------|--------|-------|--------|-------|------|--------|-------|
| Austria| 50%    | 22%       | 19%   | 24%   | 18%   | 17%    | 29%   | 24%    | 22%   | 17%    | 8%    | 34%  | 37%    | 29%   |
| Czech Republic | 51%    | 42%       | 28%   | 16%   | 9%    | 5%     | 15%   | 19%    | 23%   | 28%    | 15%   | 22%  | 43%    | 35%   |
| Denmark | 50%    | 20%       | 21%   | 22%   | 20%   | 17%    | 26%   | 20%    | 25%   | 16%    | 13%   | 37%  | 25%    | 38%   |
| France | 51%    | 23%       | 20%   | 21%   | 20%   | 15%    | 29%   | 23%    | 17%   | 23%    | 9%    | 30%  | 35%    | 36%   |
| Germany | 50%    | 20%       | 18%   | 25%   | 21%   | 16%    | 25%   | 23%    | 21%   | 20%    | 11%   | 41%  | 36%    | 24%   |
| Greece | 47%    | 33%       | 29%   | 25%   | 10%   | 3%     | 37%   | 28%    | 19%   | 11%    | 5%    | 36%  | 36%    | 28%   |
| Hungary | 51%    | 43%       | 28%   | 14%   | 8%    | 6%     | 18%   | 20%    | 15%   | 21%    | 26%   | 19%  | 48%    | 33%   |
| Italy  | 50%    | 19%       | 22%   | 23%   | 19%   | 17%    | 27%   | 26%    | 22%   | 18%    | 7%    | 35%  | 26%    | 39%   |
| Netherlands | 50%   | 22%      | 17%   | 23%   | 21%   | 17%    | 30%   | 22%    | 19%   | 19%    | 9%    | 28%  | 33%    | 39%   |
| Norway | 49%    | 22%       | 22%   | 21%   | 19%   | 16%    | 28%   | 26%    | 19%   | 13%    | 13%   | 34%  | 24%    | 42%   |
| Poland | 49%    | 19%       | 30%   | 25%   | 15%   | 11%    | 15%   | 22%    | 20%   | 24%    | 20%   | 23%  | 39%    | 38%   |
| Spain  | 50%    | 22%       | 25%   | 22%   | 17%   | 14%    | 23%   | 25%    | 16%   | 16%    | 17%   | 51%  | 23%    | 27%   |
| Sweden | 49%    | 22%       | 20%   | 21%   | 19%   | 18%    | 22%   | 15%    | 20%   | 18%    | 25%   | 37%  | 26%    | 37%   |
| Switzerland | 50%  | 22%      | 20%   | 23%   | 19%   | 16%    | 33%   | 29%    | 20%   | 14%    | 4%    | 29%  | 30%    | 41%   |
| United Kingdom | 52% | 26%       | 19%   | 21%   | 18%   | 16%    | 29%   | 21%    | 19%   | 19%    | 12%   | 27%  | 43%    | 30%   |
Table B6: Summary statistics, continued

“Lower Secondary and Below” corresponds to EISCED 1-2, “Upper Secondary” EISCED 3-4, “Advanced Vocational” EISCED 5, “BA Level” EISCED 6, and “MA Level and Above” EISCED 7.

| Education | Lower Secondary and Below | Upper Secondary | Advanced Vocational | BA Level | MA Level and Above |
|-----------|---------------------------|-----------------|---------------------|----------|--------------------|
| Austria   | 8%                        | 53%             | 20%                 | 5%       | 13%                |
| Czech Republic | 5%             | 56%             | 15%                 | 10%      | 14%                |
| Denmark  | 22%                       | 37%             | 10%                 | 22%      | 8%                 |
| France   | 10%                       | 45%             | 21%                 | 9%       | 16%                |
| Germany  | 5%                        | 38%             | 29%                 | 15%      | 13%                |
| Greece   | 4%                        | 32%             | 19%                 | 29%      | 16%                |
| Hungary  | 4%                        | 49%             | 18%                 | 19%      | 9%                 |
| Italy    | 13%                       | 48%             | 4%                  | 11%      | 23%                |
| Netherlands | 32%              | 34%             | 7%                  | 12%      | 15%                |
| Norway   | 13%                       | 35%             | 17%                 | 23%      | 12%                |
| Poland   | 9%                        | 35%             | 13%                 | 14%      | 29%                |
| Spain    | 23%                       | 19%             | 16%                 | 17%      | 25%                |
| Sweden   | 12%                       | 38%             | 28%                 | 11%      | 11%                |
| Switzerland | 10%            | 50%             | 24%                 | 7%       | 9%                 |
| United Kingdom | 17%       | 37%             | 14%                 | 21%      | 11%                |
APPENDIX C: ADDITIONAL APPLICATIONS

A. Framing Experiment on Support for Public Assistance

Many scholars of American voter behavior and political psychology study how issue frames can be used by politicians to influence public attitudes, either positively or negatively, toward specific policies. One of the simplest yet most powerful and prominent examples is the framing of policies on public assistance to the poor as “welfare.” While the term welfare is simply another label for public assistance to the poor, scholars have long noted that substantial segments of the U.S. voting population hold a peculiar aversion toward this label (Kluegel and Smith 1986; Smith 1987; Rasinski 1989; Shaw and Shapiro 2002). Furthermore, years of concrete empirical evidence for this phenomenon have been compiled, after the General Social Survey (GSS) beginning in the 1980s introduced a question-wording experiment that measures the effect of using the label “welfare” rather than “assistance to the poor” on support for assistance to the poor. As the data from this experiment across years of GSS waves have shown, referring to welfare causes enormous increases—on the order of 30-50 percentage points—in the probability of a respondent asserting that “too much” is being spent on such policies in the United States.

The experimental evidence in the GSS has led to an abundance of subsequent scholarship focused on explaining the underlying sources of this “welfare-label effect” and variation in its magnitude across different subsets of voters. In one example, Green and Kern (2012) develop a method of applying Bayesian Additive Regression Trees (BART) to model and discover treatment effect heterogeneity and apply the method to the GSS welfare framing experiment data. In their study, they find that substantial amounts of heterogeneity in the welfare-label effect manifest as a function of a number of covariates, including party identification, with the welfare-label effect being more pronounced among Republicans than Democrats. Green and Kern’s study is an example of the recent research trend in the methodological literature on integrating modern machine-learning techniques into the goal of investigating heterogeneous treatment effects. Yet while discovering the existence and extent of treatment effect heterogeneity may often be a researcher’s goal, these techniques are not designed to estimate the causal moderation effects of interest in this study.

For instance, Green and Kern’s finding that the welfare-label effect magnitude depends upon party identification does not, nor do Green and Kern claim it to, demonstrate that party identification itself is causing variation in the welfare effect. There are, of course, reasons to believe that party identification may moderate the welfare effect to some extent. A voter’s party identification relates directly to how responsive that voter is to party cues, and party cues involve delivery of partisan ideological messages, position preferences, and...
issue frames. Given the economically conservative preference for limited redistribution of the Republican party, it is reasonable to expect that Republican party cues to voters include calls for limiting social assistance, and that such calls might frame social assistance using the stigmatized label of welfare. For these reasons, we might expect that identification with the Republican party would, in fact, make an individual more sensitive to the welfare frame and hence exhibit a stronger welfare-label effect than had s/he not adopted that party identification.

In contrast, however, there are many non-political factors known to influence voters’ party identification choices that may also make voters more sensitive to the welfare frame for various reasons not directly related to party identification itself. For instance, some researchers have explored the possibility of interactions between the welfare-label effect and such variables as education, racial perceptions, and attributional tendencies [Federico 2004, Henry et al. 2004]. Such variables may also correlate with party identification such that any conditionality in the welfare-label effect across voters with different party identifications may be at least partially the result of these other variables.

To investigate the causal moderation of the welfare effect attributable to party identification, the parallel estimation framework is applied to the GSS data, and estimates of the causal moderation effect using parallel regression and parallel matching are compared to moderation effect estimates using conventional methods. As described above, the randomized treatment is whether the respondent was asked about “welfare” (coded as 1) or “assistance to the poor” (coded as 0), and the outcome is whether the respondent stated that “too much” is being spent on such policies (coded as 1) or that “too little”/“about the right amount” is being spent on such policies (coded as 0). The moderator is a dichotomous version of party identification, with those respondents identifying as Republicans coded as 1 and all others coded as 0.\footnote{In the GSS survey, respondents were asked “Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or what?” Respondents who chose the options “Strong Republican” and “Not very strong Republican” are coded as 1 here, and all other respondents are coded as 0.}

Covariates used as controls variables include age, sex, race, highest education level, real income, number of children, region, city/town type (based on size and urbanicity), year surveyed, and racial attitudes. While scholars of American voting behavior have long considered racial attitudes to be possible determinants of party identification (Carmines and Stimson 1989), some studies have cast doubt on the precise causal directionality between racial attitudes and party identification (e.g. Abramowitz 1994). In reality, the two are likely co-determined, but for the purposes of this investigation, it is illuminating to isolate the party identification moderation effect that is not related to racial attitudes. For this reason, racial attitudes are included as a control variable, with the coding...
of this variable following that used by Green and Kern (2012).

Figure C1 presents the results, with point estimates and 95% confidence intervals estimated using the conventional approaches and parallel estimation framework respectively. As

![Figure C1: Moderation of Welfare Framing Effect by Partisanship](image)

illustrated in the upper half of the plot, the results from the conventional estimators suggest substantial moderation. The first estimator is the difference in subset effects, where the two subsets are Republican respondents and non-Republican respondents, and each subset effect is the simple difference between the probability of believing that too much is being spent on poverty assistance in the treatment condition (i.e. given the “welfare” frame) and the probability of believing that too much is being spent on poverty assistance in the control condition (i.e. given the “assistance to the poor” frame). The difference in subset effects highlights important heterogeneity of the welfare effect: while there is a welfare effect among all partisan groups, the magnitude of the estimated welfare effect is almost 10 percentage points higher among Republicans than among other respondents.\(^{12}\)

The second estimator whose results are presented in Figure C1 is the controlled interaction regression estimator, which is the estimate of the interaction effect between the treatment and moderator in a regression of the outcome on the treatment, moderator, treatment-moderator interaction, and control variables. While a number of the control variables are correlated with party identification (as should be expected), and while we should also expect a number of these control variables to bear upon voters’ susceptibility to the welfare effect,

\(^{12}\)The estimated effect among non-Republicans is 0.329, while it grows over 25% to 0.418 for Republicans.
the figure shows that the inclusion of controls in the interaction regression has virtually no effect on the moderation effect estimate, compared to the simple difference in subset effects. This may seem surprising, unless one takes into account the identification result and corresponding estimation strategies presented in this study, and considers that the interaction regression does not include interactions between the treatment and controls.

As explained in this study, it would be incorrect to interpret the upper two estimates in Figure C1 as causal moderation effects because they are based on estimators that do not perform covariate adjustment in a way that allows for estimation of an identified causal moderation estimand. In contrast, results from estimation using the parallel regression method and the parallel matching method are reported in the lower half of Figure C1. As can be seen, these results paint a much more conservative picture in terms of the extent to which party identification *per se* moderates the welfare effect. Specifically, the causal moderation effect estimates yielded by these methods are attenuated toward zero and cut roughly in half.

These results indicate that a substantial portion of the moderation effect attributed to party identification by the conventional estimators is actually the result of party identification being correlated with other variables that themselves moderate the welfare effect. In particular, results from the parallel regression method include statistically significant interactions between the treatment and a number of other variables—including education level, real income, city/town type, and racial attitudes—that are also correlated with party identification.

---

13The matching procedure used Mahalanobis distance and matched all of the moderated (Republican) units with the closest unmoderated units, with replacement. Hence the estimates should be interpreted as the ATMEs for the Republicans.
B. Field Experiment on Corruption

Recent research in economics and political science has increasingly used experimental designs to study the effectiveness of various policies to curb undesirable behavior by elites and politicians, such as engagement in corruption and clientelism. One well-known example is a study by Olken (2007) on the effects of auditing on corruption in construction projects in Indonesia. In one of his analyses, Olken investigates how having one’s village construction project be randomly assigned to an auditing treatment affected a household’s probability of having a member hired to work on the project. While Olken shows (in a separate analysis) that the audit treatment decreases the misappropriation of project funds, he theorizes that this decrease in one type of corruption may be accompanied by the increase in a different type, namely nepotism. To investigate this, he tests whether having a family member as the construction project head moderates (namely, increases) the audit treatment effect on the probability of having a household member hired to work on the project.

As is standard in the literature, Olken employs a controlled interaction regression to estimate this moderation effect, regressing the outcome on the treatment, moderator, and treatment-moderator interaction, as well as including a wide range of control variables but not interacting them with the treatment. The estimated interaction in this specification between the audit treatment dummy and a dummy indicator for having a family member as the construction project head is 0.138, suggesting that having a family member as the construction project head increases the effect of the audit treatment on the outcome by approximately 14 percentage points.\textsuperscript{14} If the parallel regression approach presented in this study (or equivalently, a single regression including interactions between the treatment and all possible confounding variables) is used, the moderation effect is attenuated to 0.120, but remains statistically significant.

\textsuperscript{14}Olken also includes a specification that includes interactions between the treatment and two other possible moderator variables in addition to the interaction discussed here, but it does not include interactions between the treatment and the full range of potential confounding variables.
Appendix D: Sensitivity Analysis

This appendix shows how sensitivity analysis can be applied to estimation of the ATME using the parallel estimation framework. More specifically, it describes an adaptation of a sensitivity analysis developed by Imbens (2003). Further, for illustrative purposes, it applies the adapted sensitivity analysis to the auditing and corruption field experiment (Olken, 2007) described in Appendix C.

Imbens’ sensitivity analysis in its original form involves a parametric model in which several simplifying modeling assumptions are made. First, the outcome variable $Y$ is assumed to follow a normal distribution with a mean that is a function of a non-exogenous binary treatment $T$, observed confounders $X$, and an unobserved confounder $U$. Second, given $U$ and $X$, the treatment $T$ is modeled as following a logistic distribution. Third, the unobserved confounder $U$ is modeled as a Bernoulli random variable that is, without loss of generality, independent of $X$. By making these modeling assumptions, Imbens specifies a likelihood function whereby maximum likelihood estimation (MLE) can be used to obtain the hypothetical estimates of the treatment effect had the hypothetical confounder $U$ been observed. These estimates depend upon specified relationships, in the form of parameter values fixed by the analyst, between $U$ and $Y$ (effect of $U$ on $Y$) and between $U$ and $T$ (selection effect into $T$ given $U$). Using subject matter expertise combined with parameter estimates involving observed confounders, the analyst can choose plausible ranges of parameters to model the confounding by $U$ and, under those hypothetical conditions, estimate the extent to which the estimated treatment effect would change—i.e. its sensitivity to unobserved confounding.

The version of this sensitivity analysis extended from the ATE to ATME context, developed and presented in this study, proceeds in the same manner with two major differences. First, instead of including the treatment in the parametric model described above, the moderator is included in its place. This is because the scenario to which the parallel estimation framework pertains is one in which the treatment is randomized/exogenous but the moderator is not. Thus, it is potential unobserved confounding of the moderator, not the treatment, that must be modeled. Second, in accordance with the parallel estimation process, the full parametric model and subsequent estimation via MLE is performed separately for the treatment and control subsets.

Formally, for treatment $T_i$, moderator $S_i$, outcome $Y_i$, control variable vector $X_i$, and hypothetical unobserved confounder $U_i$, the following model is applied separately for subsets
$T_i = 0, 1$: 

$$U_i \sim B(1, 1/2)$$

$$Pr(S_i = 1|X_i, U_i) = \frac{\exp(\zeta + X_i\eta + \alpha U_i)}{1 + \exp(\zeta + X_i\eta + \alpha U_i)}$$

$$Y_i(S_i)|X_i, U_i \sim N(\xi + \gamma S_i + X_i\beta + \kappa U_i, \sigma^2)$$

As in the standard version of the Imbens sensitivity analysis, the parameters relating the hypothetical unobserved confounder $U$ with the moderator $S$ and outcome $Y$ ($\alpha$ and $\kappa$ respectively) are fixed to certain plausible values specified by the analyst in accordance with their (possibly worst-case) expectations. In the standard version of the Imbens sensitivity analysis, because the model is applied to only one sample of data, only two parameters must be specified. In contrast, in the version presented here, there are four parameters to be fixed since there is an $\alpha$ and $\kappa$ for each treatment subset.

The enlargement of the fixed parameter set may, at first consideration, appear to overcomplicate this version of the sensitivity analysis. However, this turns out not to be the case. First, given Assumptions 2 and 3 that the treatment is randomized and that the moderator and confounding variables (including $U$) are not affected by the treatment, the selection effect into $S$ given $U$ (i.e. $\alpha$) is equal for both the treatment and control subsets. Hence, when fixing the $\alpha \equiv \tilde{\alpha}$ values in the sensitivity analysis, $\tilde{\alpha}$ should be set to the same value for both subsets $T_i = 0, 1$, reducing as before to a single fixed parameter. In contrast, the same cannot be done for $\kappa$, the effect of $U$ on $Y$. For there to be moderation of the treatment effect by $S$, then it is also the case that $S$ affects $Y$ differently depending upon whether $T_i = 0$ or $T_i = 1$. Similarly, the confounding of the estimate of the ATME by an unobserved variable is due precisely to the possibility that the unobserved confounding variable also affects $Y$ differently depending upon whether $T_i = 0$ or $T_i = 1$. In other words, we are concerned that the ATME estimate is confounded because an unobserved variable also moderates the treatment effect. Further, just as $\gamma_{T_i=1} - \gamma_{T_i=0}$ represents the ATME of interest, $\kappa_{T_i=1} - \kappa_{T_i=0}$ represents the analogous moderation of the treatment effect by $U$. As a result, while $\kappa_{T_i=1}$ and $\kappa_{T_i=0}$ must be fixed separately, the sensitivity analysis should focus on their difference, reducing the task of interpretation, if not the underlying parameters, to a single value of interest.

In sum, this sensitivity analysis allows for the investigation of how sensitive the ATME estimate is (e.g. the extent to which it would be attenuated toward zero) given two primary considerations: a fixed hypothetical selection effect into the moderator given $U$ and a fixed hypothetical moderation effect on the treatment caused by $U$. Formally, for fixed $\tilde{\alpha}$ and...
\( \tilde{\kappa}_{T_i=1} - \tilde{\kappa}_{T_i=0}, \) the implied estimate of the ATME, \( \tilde{\delta} = \tilde{\gamma}_{T_i=1} - \tilde{\gamma}_{T_i=0}, \) can be computed via MLE.

To illustrate, the adapted sensitivity analysis is applied to the Olken (2007) experimental data on the effects of auditing on corruption in construction projects in Indonesia. As described earlier, Olken tests whether having a family member as the head of the construction project moderates (namely, increases) the audit treatment effect on the probability of having a household member hired to work on the project. While this turns out to be a case in which specifying the proper model for estimating causal moderation does not dramatically change the resulting estimate relative to conventional approaches, there is still the possibility of the ATME estimate being confounded by unobserved variables. Thus, the sensitivity analysis as described above is applied to the Olken data and model\(^{15}\). To enable easier inspection and visualization of the sensitivity of the causal moderation effect, the implied \( \tilde{\delta} \) was estimated over a grid of plausible \( \tilde{\alpha} \) and \( \tilde{\kappa} \) values. Figure D1 displays a set of results from the sensitivity analysis.

Under consideration in the figure is what strength of unobserved confounding would be required to cut the estimated ATME in half (though any value of interest could be specified for this). The \( y \)-axis measures the hypothetical selection effect into the moderator attributable to the unobserved confounder (\( \tilde{\alpha} \)), while the \( x \)-axis measures the hypothetical causal moderation of the treatment by the unobserved confounder (\( \tilde{\kappa}_{T_i=1} - \tilde{\kappa}_{T_i=0} \)). The blue curve is a level curve depicting the combination of values for these two effects required to halve the estimated ATME. The vertical red and horizontal green dashed lines provide reference points for interpreting the extent of confounding implied by the level curves. The horizontal green dashed line marks the largest (in absolute value) estimated selection effect by a variable that is observed in the data. This variable is closely related to the moderator (whereas the moderator is having a family member as project head, this other variable is an indicator for having a family member in the village government). Further, it is also a binary variable, like the unobserved confounder, thus providing a clear theoretical reference point for how strong of a relationship is implied by \( \tilde{\alpha} \) values above the green line. The vertical red dashed line marks the parallel regression estimate of the ATME, thus providing a clear reference point for the hypothetical causal moderation effect of the unobserved confounder.

In sum, if the true ATME is only half the size of the estimated ATME due to estimation bias caused by an unobserved confounder, the blue level curve depicts how significant that confounder would need to be in terms of its relationship with the moderator of interest and

\(^{15}\)For computational efficiency purposes, the stratum fixed effects included in Olken’s model are dropped. These stratum fixed effects have little effect on the estimates in the original models, thus suggesting that dropping them is unlikely to seriously affect the inferences drawn from this analysis.
its own impact on the treatment effect. Because the blue level curve does not pass through the lower-left box formed by the red and green lines—which may be considered a sensitivity danger zone—the results of this sensitivity analysis can be interpreted as evidence that it is unlikely for there to exist an unobserved confounder significant enough to imply an ATME that is only half the size of the estimated ATME. Thus, while the existence of unobserved confounders is virtually certain, we can be relatively confident in the existence of a causal moderation effect of the sort posited by Olken.