# Quantifying Robustness to External Validity Bias*

Martin Devaux† Naoki Egami‡

First Version: April 11, 2022
This Version: September 7, 2022

**Abstract**

The external validity of experimental results is essential in the social sciences. Existing methods estimate causal effects in a target population, called the target population average treatment effect (T-PATE). However, these methods are sometimes difficult to implement either because it is infeasible to obtain data for the target population or because there is no target population that analysts and skeptics can agree on. We consider a different goal — quantifying how robust an experiment is to external validity bias. In particular, we propose a measure of *external robustness* by estimating how much different a population should be from the experimental sample to explain away the T-PATE. Large estimated external robustness implies that causal conclusions remain the same unless populations of interest are significantly different from the experimental sample. Unlike the standard generalization approach, estimation of external robustness only requires experimental data and does not require any population data. We prove that the proposed estimator is consistent to the true external robustness under common generalization assumptions and, more importantly, has a simple interpretation even when those assumptions are violated. We provide benchmarks to help interpret the degree of external robustness in each application.

*Keywords:* Causal inference, External validity, Generalization

---

*The proposed methodology is implemented via our open-source software R package, `exr`, available at `https://github.com/naoki-egami/exr`. We would like to thank Avidit Acharya, Matt Blackwell, Neal Beck, Alex Coppock, Christian Fong, Don Green, Jens Hainmueller, Erin Hartman, Dan Hopkins, Melody Huang, Macartan Humphreys, Kosuke Imai, Gary King, Shiro Kuriwaki, Walter Mebane, John Marshall, Cyrus Samii, Tara Slough, Brandon Stewart, Mike Tomz, Matt Tyler, Yiqing Xu, Teppei Yamamoto, and Soichiro Yamauchi for their thoughtful comments. We would like to especially thank Dominik Rothenhäusler for insightful discussions on this paper and other problems on external validity. Finally, we appreciate comments from participants at a seminar at Stanford and the 2022 Political Methodology meeting.

†Ph.D. student, Department of Political Science, Columbia University, New York, NY 10027. Email: msd2202@columbia.edu, URL: `https://www.martindevaux.com`

‡Assistant Professor, Department of Political Science, Columbia University, New York, NY 10027. Email: naoki.egami@columbia.edu, URL: `https://naokiegami.com`

# 1 Introduction

Over the last couple of decades, social scientists have paid greater attention to issues of causal inference and applied a host of experimental and quasi-experimental methods. The central focus of this trend has been on *internal validity* — researchers try to unbiasedly estimate causal effects within a study, without making strong assumptions. The *external validity* of randomized experiments has been one of the most important long-standing methodological questions in the social sciences (e.g., Shadish, Cook and Campbell, 2002; Bareinboim and Pearl, 2016; Findley, Kikuta and Denly, 2020; Wilke and Humphreys, 2020; Egami and Hartman, 2022).

One central question of external validity is whether causal estimates in an experiment can be generalized to other relevant populations. To answer this question, the large literature on generalization has developed various approaches, including weighting methods popular in the social sciences, to estimate causal effects in a target population, which is known as the target population average treatment effect (T-PATE). These methods require not only experimental data but also separate data for the target population (e.g., data from a nationally representative survey or from the census). At its core, they approximate the target population via weighting and/or regression approaches (e.g., Imai, King and Stuart, 2008; Gerber and Green, 2012; Tipton, 2013; Miratrix et al., 2018; Dahabreh et al., 2019; Egami and Hartman, 2021).

Despite the importance of these recent methodologies, they are sometimes difficult to implement in practice for several reasons. First, multiple relevant target populations can exist, and it is often difficult to justify one particular choice of the target population in practice. Skeptics might be interested in whether experimental results generalize to other plausible populations. When there is no one target population that both analysts and skeptics can agree on, researchers are often reluctant to estimate the T-PATE for any population, and they conduct no empirical evaluation of external validity. Second, even if researchers are willing to specify one particular target population, it is often difficult to obtain data for the target population. In some settings, researchers are unable to collect any data. In other settings, researchers can collect some basic demographic variables (e.g., gender, age, race, and education) for the target population, and yet, they cannot measure a richer set of covariates that are likely to be important for treatment effect heterogeneity and selection into the experiment. For these practical reasons, even when analysts are aware of severe concerns about external validity, they might not be able to estimate the T-PATE and empirically evaluate external validity.[1]

[1]Egami and Hartman (2022) find that only 11% of all experimental studies published in the American Political Science Review from 2015 to 2019 contain a formal analysis of external

In this paper, we consider a different goal — quantifying how robust an experiment is to external validity bias. In particular, we propose a simple measure of *external robustness*, which ranges from 0 to 1, by estimating how much different a population should be from the experimental sample to explain away the T-PATE (Section 3). If estimated external robustness is large, this means stronger robustness to external validity bias; causal conclusions in an experiment stay the same unless populations of interest are significantly different from the experimental sample. On the other hand, smaller estimated external robustness implies that causal conclusions in an experiment can change even in populations that are only slightly different from the experimental sample. This is a transparent way to reveal robustness to external validity bias. We formally answer this question by estimating the amount of reweighting of the experimental data required to make the T-PATE estimate equal to zero.

The proposed approaches have three desirable properties. First, unlike the standard approach for generalization, estimation of external robustness only requires experimental data. That is, researchers do not need to specify/justify a particular target population or collect population data. This is because our goal is not to estimate the T-PATE but to characterize robustness to external validity bias. In practice, researchers can estimate external robustness in any experimental study without collecting additional data.

Second, we prove that the proposed estimator for external robustness is consistent to the true external robustness under common generalization assumptions, known as the ignorability of sampling and treatment effect heterogeneity assumption and the overlap assumption (Egami and Hartman, 2022). More importantly, we also show that the proposed estimator is consistent to the upper bound of the true external robustness even when those generalization assumptions are violated. Thus, regardless of whether those assumptions hold, estimated external robustness has a simple interpretation; low estimated external robustness reveals the lack of robustness to external validity bias. While high estimated external robustness does not guarantee high external validity, it is necessary for external validity. We also show how researchers can use an explicit sensitivity analysis to investigate how close the upper bound is to the true external robustness.

Finally, we provide simple default benchmarks to substantively interpret whether estimated external robustness is "large" or "small" in each application (Section 4). We clarify how we can rely on national surveys (e.g., ANES and CES) and online samples (e.g., Amazon Mechanical Turk samples) to construct benchmarks that can be used in any application.

---

validity in the main text. Findley, Kikuta and Denly (2020) also find that only exceptional few papers contained a dedicated external validity discussion.

In sum, our paper requires less data and weaker assumptions because our goal of estimating external robustness is more modest than the goal of traditional generalization methods to estimate the T-PATE. We clarify that, while our approach cannot directly estimate the T-PATE, a measure of external robustness provides a transparent way to reveal robustness to external validity bias. Thus, this paper provides a middle ground approach — assessing external validity under assumptions plausible in practice, using only experimental data that researchers already have — in contrast to traditional methods that aim to estimate the T-PATE, which can be too challenging because they require more stringent assumptions and population data that are difficult to obtain.

To clarify practical considerations, we summarize when to use and how to use the proposed method in Section 5. Importantly, all of the proposed methods can be implemented via our companion `R` package `exr`. We also clarify common concerns and potential limitations. We illustrate our proposed approach in Section 6 using a field experiment by Domurat, Menashe and Yin (2021). In the Appendix, we provide another empirical application based on a survey experiment (Appendix D), and we empirically validate the performance of external robustness using replication experiments collected by Coppock, Leeper and Mullinix (2018) (Appendix F). Finally, in the Appendix, we also discuss several key extensions of the proposed method (including external validity with respect to contexts, and observational studies) as well as proofs of theoretical results and simulation studies.

### Related Literature

Most existing methods of generalization and external validity consider how to estimate the T-PATE for a particular target population (e.g., Imai, King and Stuart, 2008; Cole and Stuart, 2010; Hartman et al., 2015; Kern et al., 2016). The main difference from these methods is that our approach asks the distinct question of external robustness, and thus, most importantly, we do not require any population data, which is the key practical challenge we tackle in this paper.

There are several lines of research that address goals similar to ours. First, Stuart et al. (2011) and Tipton (2014) consider how to assess the generalizability of an experimental sample for a particular target population. Like these papers, we focus on assessing robustness to external validity bias. However, unlike ours, these papers require a particular target population data and assess generalizability against the selected target population, which is a separate goal. Second, the question of external robustness is similar in spirit to sensitivity analyses for causal inference, which ask how robust causal conclusions are to unmeasured confounding. In the literature on generalization, recent papers develop sensitivity analyses considering the violation of the ignorability of sampling and treatment effect heterogeneity assumption (e.g., Nguyen

et al., 2017; Andrews and Oster, 2019; Nie, Imbens and Wager, 2021; Huang, 2022). Unlike ours, these methods also assume data from a particular target population, while some relevant variables might be unobserved. The goal of estimating external robustness is also different from the goal of sensitivity analyses to estimate the bound of the T-PATE. Therefore, these sensitivity analyses are complementary to our approach. Indeed, when we prove consistency of our estimator to the upper bound in cases where common generalization assumptions are violated, we incorporate theoretical results in sensitivity analyses (see Section 3.3.2).

Our formalization builds on a long tradition of the entropy optimization problem. In particular, our paper is closest to an important recent work by Gupta and Rothenhäusler (2021). In the context of the distributional shift literature, they propose the $s$-value, which quantifies robustness to the change in covariate distributions. This contains the generalization of randomized experiments as a special case. Their work focuses on the properties of $s$-values in a general setting, and is complementary to our work focusing on external validity in the potential outcomes framework. Spini (2021) also considers the similar question of robustness to distributional shift. Their work casts the estimation of robustness as the debiased-GMM problem under the common assumptions of generalization. In this paper, we make a number of unique contributions that are practically important; we explicitly consider the violation of the common generalization assumptions and prove consistency to the upper bound in such settings (Section 3), we propose interpretable default benchmarks that applied researchers can use in any application (Section 4), we summarize detailed practical consideration (Section 5), and we provide empirical validation studies using replication experiments (Section 6 and Appendix F).

Finally, we primarily focus on external validity in terms of populations, which is called $X$-validity (Egami and Hartman, 2022), because it is one of the most common concerns we face in practice given that many experiments use some forms of convenience or non-probability samples. However, we also acknowledge that there are other important dimensions of external validity, in particular, external validity with respect to contexts (e.g., Bareinboim and Pearl, 2016; Deaton and Cartwright, 2018; Munger, 2019; Blair and McClendon, 2020; Egami and Hartman, 2022). To address this issue, researchers typically rely on a method complementary to our approach, i.e., meta-analysis based on multiple experiments or multi-site experiments. When researchers have many comparable studies that allow for clear meta-analyses, a variety of methods are available to aggregate causal effects (e.g., Cooper, Hedges and Valentine, 2019; Dunning et al., 2019). However, in practice, it is rare to have many scientific replications of the same experiment, and Slough and Tyson (2022) recently show that it is more difficult to have "target-equivalent" studies than previously thought. Our paper considers a distinct goal — allow

each experimental study to assess the robustness of its causal conclusions to external validity bias. In Appendix B, we also show that researchers can estimate external robustness in terms of contexts using a single experimental data in certain settings when contextual moderators can explain the mechanism behind contextual differences. As the issue of external validity is complex and challenging, we believe both complementary approaches are useful.

# 2 Setup and Methodological Challenges

## 2.1 Setup

Consider a randomized experiment with a total of $n$ units, each indexed by $i \in \{1, \ldots, n\}$. We use $\mathcal{P}_{\texttt{exp}}$ to denote a distribution from which the experimental samples were drawn. Within the randomized experiment, a treatment variable $T_i$ is randomly assigned to each respondent, potentially conditional on some covariates (e.g., block randomization). For notational simplicity, we focus on a binary treatment $T_i \in \{0, 1\}$, but the same framework is applicable to categorical and continuous treatments with appropriate notational changes. Using the potential outcomes framework, we then define $Y_i(t)$ to be the potential outcome of unit $i$ if the unit were to receive the treatment $T_i = t$ where $t \in \{0, 1\}$.

When researchers randomize treatments in an experiment, we can use simple estimators, such as difference-in-means, to estimate the *sample average treatment effect* (SATE).

$$\mathsf{SATE} \coloneqq \mathbb{E}\{Y_i(1) - Y_i(0); \mathcal{P}_{\texttt{exp}}\}, \tag{1}$$

where the expectation is taken over $\mathcal{P}_{\texttt{exp}}$. The SATE represents the causal effect in the experimental sample distribution $\mathcal{P}_{\texttt{exp}}$.

The main issue of external validity is that researchers are not only interested in this within-experiment estimand but also interested in whether causal conclusions are generalizable to other relevant populations.

## 2.2 Standard Approach for Generalization: Review

The standard approach for generalization considers how to estimate a causal effect in a target population data, which differs from the experimental sample (e.g., Cole and Stuart, 2010; Tipton, 2013; Hartman et al., 2015; Kern et al., 2016; Dahabreh et al., 2019). This approach requires that researchers specify a concrete target population $\mathcal{P}^*$ and collect its data. Formally, the *target population average treatment effect* (T-PATE) is defined as,

$$\mathsf{T\text{-}PATE}(\mathcal{P}^*) \coloneqq \mathbb{E}\{Y_i(1) - Y_i(0); \mathcal{P}^*\}, \tag{2}$$

where the expectation is taken over $\mathcal{P}^*$ instead of $\mathcal{P}_{\texttt{exp}}$ in the SATE.

Identification of the T-PATE relies on two assumptions. The first assumption requires that selection into the experiment and treatment effect heterogeneity be unrelated to each other after adjusting for pre-treatment covariates $\mathbf{X}_i$ (Cole and Stuart, 2010).

**Assumption 1 (Ignorability of Sampling and Treatment Effect Heterogeneity)** *For all* $\mathbf{x} \in \mathcal{X}$ *where* $\mathcal{X}$ *is the support of* $\mathbf{X}_i$,

$$\mathbb{E}(Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}; \mathcal{P}_{exp}) = \mathbb{E}(Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}; \mathcal{P}^*). \tag{3}$$

This assumption states that the conditional average treatment effects (CATEs) are the same in the experimental sample and target population. This is the assumption researchers make when they use sampling weights to adjust for the representativeness of survey or experimental data. In a special case of random sampling, this assumption holds without any covariate adjustment.

The second key assumption for generalization is that all types of units in the target population $\mathcal{P}^*$ can be selected into the experiment. Intuitively, if there are types of units in the target population that are not represented at all in the experimental data, a randomized experiment cannot help researchers learn causal effects on them, and thus, generalization is impossible. This assumption is violated when, for example, the target population data includes people aged 65 years and older, and yet, the experimental data does not contain any observation for this subset.

**Assumption 2 (Overlap)** *For all* $\mathbf{x} \in \mathcal{X}$,

$$\Pr(\mathbf{X}_i = \mathbf{x}; \mathcal{P}^*) > 0 \Longrightarrow \Pr(\mathbf{X}_i = \mathbf{x}; \mathcal{P}_{exp}) > 0, \tag{4}$$

*where* $\Pr(\mathbf{X}_i = \mathbf{x}; \mathcal{P}^*)$ *and* $\Pr(\mathbf{X}_i = \mathbf{x}; \mathcal{P}_{exp})$ *represent the probabilities of* $\mathbf{X}_i = \mathbf{x}$ *in the target population* $\mathcal{P}^*$ *and experimental sample* $\mathcal{P}_{exp}$.

For notational simplicity, this paper uses probability notations to denote the distribution of observed covariates $\mathbf{X}_i$, but the same results follow when we write them with more general probability density functions.

Given a randomized experiment, under Assumptions 1 and 2, we can identify the T-PATE as follows.

$$\mathsf{T\text{-}PATE}(\mathcal{P}^*) = \sum_{\mathbf{x} \in \mathcal{X}} \underbrace{\tau_0(\mathbf{X}_i = \mathbf{x})}_{\substack{\text{CATE} \\ \text{for units with } \mathbf{x}}} \times \underbrace{\Pr(\mathbf{X}_i = \mathbf{x}; \mathcal{P}^*)}_{\substack{\text{Distribution of } \mathbf{X} \\ \text{in the target population}}},$$

where $\tau_0(\mathbf{X}_i = \mathbf{x})$ is the CATE for units with $\mathbf{x}$, and defined as,

$$\tau_0(\mathbf{X}_i = \mathbf{x}) \coloneqq \mathbb{E}(Y_i \mid T_i = 1, \mathbf{X}_i = \mathbf{x}; \mathcal{P}_{\mathtt{exp}}) - \mathbb{E}(Y_i \mid T_i = 0, \mathbf{X}_i = \mathbf{x}; \mathcal{P}_{\mathtt{exp}}).$$

The essential idea of this identification formula is that we first estimate the CATE $\tau_0(\mathbf{X}_i = \mathbf{x})$ within the randomized experiment, and then we reweight them to the distribution of $\mathbf{X}$ in the target population $\Pr(\mathbf{X}_i = \mathbf{x}; \mathcal{P}^*)$.

Using this identification formula, researchers can use a generalized weighting estimator to estimate the T-PATE.

$$\widehat{\mathsf{T\text{-}PATE}}(\mathcal{P}^*) = \frac{1}{n} \sum_{i=1}^{n} w_i \widehat{\tau}(\mathbf{X}_i) \tag{5}$$

where $\widehat{\tau}(\mathbf{X}_i)$ is an estimated CATE, and weights $w_i$ capture the difference in the distribution of $\mathbf{X}$ in the experimental sample $\mathcal{P}_{\mathsf{exp}}$ and the target population $\mathcal{P}^*$.

When researchers obtain data for the target population, they can estimate weights $w$. Researchers can model the sampling process (e.g., Cole and Stuart, 2010; Tipton, 2013) or use balancing methods, such as calibration and entropy balancing (e.g., Deville and Särndal, 1992; Hainmueller, 2012; Hartman et al., 2015). However, when researchers do not have data for the target population, they cannot estimate weights, and thus, estimation of the T-PATE is infeasible, which we elaborate on in the next section.

### 2.3 Methodological Challenges

While the standard approach for estimating the T-PATE is important, it is sometimes difficult to implement in practice for two reasons. First, there are potentially many relevant target populations, and it is often difficult to justify one particular choice of the target population in practice. When there is no one target population that both analysts and skeptics can agree on, researchers are often reluctant to estimate the T-PATE for any target population, and they conduct no formal analysis of external validity.

Second, even if researchers are willing to specify one particular target population, it is often difficult to obtain rich data for the target population. In some settings, researchers are unable to collect any data. In other settings, researchers can collect some basic demographic variables (e.g., gender, age, race, and education) for the target population, and yet, they cannot measure a richer set of covariates that are likely to be important in explaining treatment effect heterogeneity and selection into the experiment. This is a common scenario because the experimental data are collected for a specific research question that analysts are interested in, and yet, the target population data (e.g., the census data and national surveys like ANES and CES) are often not collected for the experiment in mind.

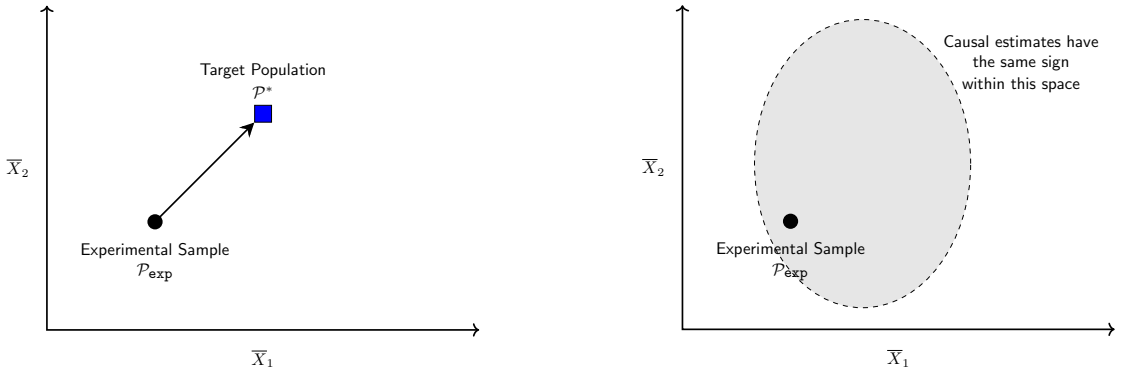
For these reasons, even when researchers are aware of acute concerns about external validity, they might not be able to estimate the T-PATE and evaluate external validity formally.

# 3 The Proposed Methodology

## 3.1 Overview

In the standard approach for generalization, researchers aim to estimate the T-PATE for a specific target population. This paper considers a different question — quantifying how robust an experiment is to external validity bias. In other words, our goal is to characterize how much different a population should be from the experimental sample to explain away the T-PATE. If we find that causal conclusions are generalizable only to populations close to the experimental sample, it implies that the experimental estimates are less robust to external validity bias. If causal conclusions are the same even in populations that are significantly different from the experimental sample, it implies stronger robustness to external validity bias.

Figure 1 visualizes the difference between our goal and the goal of the standard approach. Suppose the distribution of units can be characterized by the means of two variables $\overline{X}_1$ and $\overline{X}_2$, e.g., gender and age. A randomized experiment allows us to estimate the SATE in the experimental sample $\mathcal{P}_{\texttt{exp}}$, which is a particular point in the space. In the standard approach for generalization (see panel (a)), researchers aim to estimate the T-PATE in a specific target population $\mathcal{P}^*$, which is another point in the space. In contrast, our question of *external robustness* asks how robust experimental estimates are to different populations (see panel (b)). In other words, we estimate the size of a space within which the T-PATE estimates have the same sign as the SATE. If this space is larger (smaller), experimental results are robust to a wider (smaller) range of populations.




(a) Standard Approach:

Estimate the T-PATE for a specific target population $\mathcal{P}^*$

(b) External Robustness:

Estimate the size of a space within which the T-PATE and SATE have the same sign

Figure 1: Conceptual Overview of External Robustness.

Importantly, our approach to external robustness does not allow researchers to estimate the exact magnitude of the T-PATE. However, we can characterize the robustness of the sign of causal effects to external validity bias with much fewer data requirements and weaker assumptions, as we show in this section; (1) researchers do not need to specify/justify a particular target population or collect population data, and (2) our proposed measure of external robustness has a simple causal interpretation even when assumptions conventionally required for generalization (Assumptions 1 and 2) are violated.

### 3.2 A Measure of External Robustness

The question of external robustness is how much different a population should be from the experimental sample to explain away the T-PATE. As reviewed in Section 2.2, we can formally characterize the difference in distributions via the amount of reweighting. Therefore, the question of external robustness is formally equivalent to asking how robust experimental results are to different degrees of reweighting. If we find causal estimates are robust to a large amount of reweighting, it means that causal conclusions are robust to a wide range of populations, and external robustness is high. In contrast, if causal estimates are sensitive to a small amount of reweighting, it means that causal conclusions are only generalizable to populations that are similar to the experimental sample, and external robustness is low.

Building on this idea, we now propose a measure of external robustness. To preview one of our key contributions in Section 4, we will also provide simple default benchmarks that researchers can use to substantively interpret and justify whether estimated external robustness is large or small in any given application.

We formalize the question of external robustness in terms of a minimization problem. Without loss of generality, suppose the SATE estimate is positive. We estimate weights closest to uniform weights such that the T-PATE estimate is less than or equal to zero. Intuitively, this captures the minimum amount of reweighting required to explain away the T-PATE. This is equivalent to estimating the population that is closest to the experimental sample among populations for which the T-PATE is less than or equal to zero.

$$\begin{aligned} \widehat{\mathrm{KL}} \quad = \quad & \min_{\mathbf{w}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} w_i \log(w_i)}_{\text{KL-divergence}} \\ & \text{s.t.} \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} w_i \widehat{\tau}(\mathbf{X}_i) \leq 0,}_{\text{T-PATE constraint}} \qquad \underbrace{\sum_{i=1}^{n} w_i = n, \quad w_i \geq 0,}_{\text{Standard weights constraints}} \end{aligned} \tag{6}$$

where $\widehat{\tau}(\mathbf{X}_i)$ is an estimated CATE for units with $\mathbf{X}_i$.

This minimization problem has three components. First, the Kullback–Leibler (KL) divergence, $\frac{1}{n}\sum_{i=1}^{n} w_i \log(w_i)$, captures how much weights differ from uniform weights (i.e., weights all equal to one = no reweighting). We use the KL divergence because it is the most popular measure of the difference between two distributions (e.g., Deville and Särndal, 1992; Hainmueller, 2012).[2] This optimization finds the minimum amount of reweighting required to satisfy constraints.

The second and most important component is the T-PATE constraint $\frac{1}{n}\sum_{i=1}^{n} w_i \widehat{\tau}(\mathbf{X}_i) \leq 0$. This constraint guarantees that weights are estimated such that the resulting T-PATE estimate is less than or equal to zero. This constraint allows us to estimate the minimum amount of reweighting required to explain away the T-PATE estimate. When the SATE estimate is negative, we just need to change the T-PATE constraint to $\frac{1}{n}\sum_{i=1}^{n} w_i \widehat{\tau}(\mathbf{X}_i) \geq 0$. Here, we use zero as a threshold, but researchers can also use any other threshold to estimate external robustness such that the T-PATE is less than or equal to some substantively significant effect size (e.g., 5 percentage points). Finally, the second and third constraints are standard, and they make weights non-negative and make their sum equal to the sample size.

Then, we define the measure of *external robustness* as

$$\widehat{\xi} \coloneqq 1 - \exp(-\widehat{\mathrm{KL}}), \tag{7}$$

where $\widehat{\mathrm{KL}}$ is the minimum value of equation (6). This transformation makes the measure range from 0 to 1 (Gupta and Rothenhäusler, 2021). Small estimated external robustness implies that only a small amount of reweighting is required to explain away the T-PATE, i.e., the T-PATE is zero even in populations close to the experimental sample. Large estimated external robustness

[2]Many existing methods aim to minimize the KL divergence. For example, in non-causal survey sampling settings, Deville and Särndal (1992) develop a general class of calibration methods to estimate survey weights that approximate a chosen target population for which analysts have data. To estimate causal effects in observational studies, Hainmueller (2012) develops entropy balancing methods to estimate weights that balance treatment and control groups. These methods estimate weights by minimizing the KL divergence. However, our method and these previous methods have distinct goals, and thus, methods are different in a crucial way. These previous methods are designed for computing weights to balance covariates and estimate the T-PATE for a particular population (Hartman et al., 2015), and thus, they require data from the selected target population, unlike ours. Because our goal of estimating external robustness is distinct from this goal, our minimization problem does not require any population data and has a new T-PATE constraint.

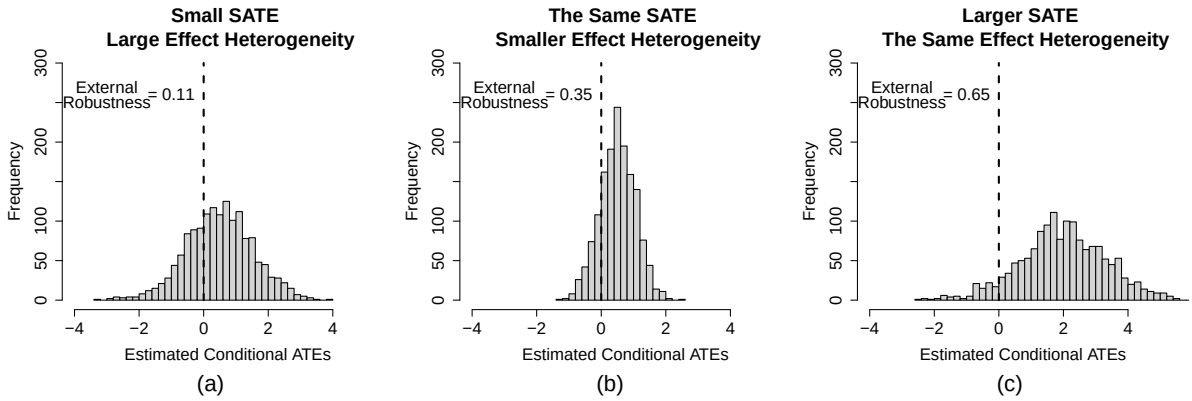



Figure 2: Illustration of External Robustness with Simulated Distributions of CATEs.

implies that a large amount of reweighting is required to explain away the T-PATE, i.e., causal conclusions remain the same even when populations are different from the experimental sample.

**Understanding External Robustness**

We now discuss how external robustness works in practice. Estimated external robustness depends on the size of treatment effects and treatment effect heterogeneity. If estimated CATEs $\widehat{\tau}(\mathbf{X}_i)$ are small, heterogeneous and range from negative to positive values, estimated external robustness is small because it is easy to satisfy the T-PATE constraint $\frac{1}{n}\sum_{i=1}^{n} w_i\widehat{\tau}(\mathbf{X}_i) \leq 0$, i.e., small reweighting is enough to make an estimate of the T-PATE equal to zero. This means that the sign of the SATE estimate and that of the T-PATE estimate will be different even if populations are only slightly different from the experimental sample.

In contrast, if estimated CATEs $\widehat{\tau}(\mathbf{X}_i)$ are large and homogeneous, estimated external robustness is large because it is harder to satisfy the T-PATE constraint $\frac{1}{n}\sum_{i=1}^{n} w_i\widehat{\tau}(\mathbf{X}_i) \leq 0$, i.e., large reweighting is required to make an estimate of the T-PATE equal to zero. This implies that the SATE estimate and the T-PATE estimate will have the same sign unless populations are significantly different from the experimental sample. In a special case when estimated CATEs are positive for all $\mathbf{x}$, there exists no weight that satisfies the T-PATE constraint, and estimated external robustness $\widehat{\xi}$ is equal to 1.

In Figure 2, we visualize how external robustness works using simple simulated data. We showcase different distributions of estimated CATEs. In Panel (a) where the SATE is small and there exists a large variation in estimated CATEs, external robustness is low ($\widehat{\xi} = 0.11$). In Panel (b), while the SATE is the same as that in Panel (a), there exists much smaller treatment effect heterogeneity, resulting in higher external robustness ($\widehat{\xi} = 0.35$). In Panel (c), while the extent of treatment effect heterogeneity is the same as that of Panel (a), the SATE is much larger, resulting in higher external robustness ($\widehat{\xi} = 0.65$).

Importantly, to estimate CATEs $\widehat{\tau}(\mathbf{X}_i)$, researchers can use a wide range of estimators in our proposed approach. While simple estimators like OLS have been conventionally used to investigate treatment effect heterogeneity, researchers can also use machine learning based estimators for the CATEs within our framework, such as causal forest (Wager and Athey, 2018), X-learners (Künzel et al., 2019), and other popular methods (e.g., Hill, 2012; Green and Kern, 2012; Imai and Ratkovic, 2013; Grimmer, Messing and Westwood, 2017; Kennedy, 2020). See Theorem 1.

**Incorporating Uncertainty**

To incorporate uncertainty, we rely on nonparametric bootstrap. Suppose the SATE estimate is positive. Then, we estimate external robustness such that the lower limit of the confidence interval of the T-PATE, approximated by nonparametric bootstrap, is equal to or smaller than zero. When the SATE estimate is negative, we estimate external robustness such that the upper limit of the confidence interval of the T-PATE is equal to or larger than zero. In practice, we recommend reporting external robustness both based on a point estimate and a bootstrap-approximated confidence interval of the T-PATE. We provide details in Appendix A.3.

### 3.3 Formal Properties

In this section, we examine the statistical properties of the proposed estimator of external robustness. We prove two main results in this section. First, for populations that satisfy common generalization assumptions (Assumptions 1 and 2), we show that the proposed estimator for external robustness is consistent to the true external robustness under mild conditions about CATE estimators. Second, and more importantly, we show that the proposed estimator for external robustness is consistent to the upper bound of the true external robustness even when those common generalization assumptions are violated.

Therefore, regardless of whether the common generalization assumptions hold, a measure of external robustness is informative about external validity; low estimated external robustness reveals the lack of robustness to external validity bias (because the true external robustness is low if the upper bound is low). While high estimated external robustness does not guarantee high external validity, it is necessary to defend its external validity. Thus, in practice, we recommend estimating and reporting external robustness, regardless of whether the common generalization assumptions hold.

### 3.3.1 Consistency

To define the true external robustness, we consider the expectation-version of the KL minimization problem.

$$\mathrm{KL}_0 \coloneqq \min_{\widetilde{\mathcal{P}} \in \mathcal{CP}} \mathrm{KL}(\widetilde{\mathcal{P}} || \mathcal{P}_{\texttt{exp}}) \qquad \text{s.t.} \quad \underbrace{\mathbb{E}\{\tau_0(\mathbf{X}_i); \widetilde{\mathcal{P}}\} \leq 0}_{\text{T-PATE constraint}} \tag{8}$$

where the KL divergence is defined as $\mathrm{KL}(\widetilde{\mathcal{P}} || \mathcal{P}_{\texttt{exp}}) \coloneqq \int \log \left( d\widetilde{\mathcal{P}} / d\mathcal{P}_{\texttt{exp}} \right) d\widetilde{\mathcal{P}}$, and $\mathcal{CP}$ is a class of *comparable* distributions $\widetilde{\mathcal{P}}$ that satisfy the following two conditions. For all $\mathbf{x} \in \mathcal{X}$,

$$\mathbb{E}(Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}; \mathcal{P}_{\texttt{exp}}) = \mathbb{E}(Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}; \widetilde{\mathcal{P}}), \tag{9}$$

$$\Pr(\mathbf{X}_i = \mathbf{x}; \widetilde{\mathcal{P}}) > 0 \Longrightarrow \Pr(\mathbf{X}_i = \mathbf{x}; \mathcal{P}_{\texttt{exp}}) > 0. \tag{10}$$

This problem aims to find comparable population $\widetilde{\mathcal{P}} \in \mathcal{CP}$ that is closest to the experimental sample $\mathcal{P}_{\texttt{exp}}$ such that the T-PATE $\mathbb{E}\{\tau_0(\mathbf{X}_i); \widetilde{\mathcal{P}}\}$ is less than or equal to zero. Therefore, the true external robustness can be defined as $\xi_0 \coloneqq 1 - \exp(-\mathrm{KL}_0)$. The minimization problem (equation (6)) we discussed in Section 3.2 solves the empirical version of this KL minimization problem.

The following theorem shows that, for comparable populations, the proposed estimator for external robustness (equation (7)) is consistent to the true external robustness under mild conditions about the CATE estimator, which allows for machine learning based estimators.

**Theorem 1** *Suppose that the CATE estimator $\widehat{\tau}$ is consistent in $L_2$, i.e., $\mathbb{E}\{(\widehat{\tau}(\mathbf{X}_i) - \tau_0(\mathbf{X}_i))^2\} \xrightarrow{p} 0$, and $|\widehat{\tau}(\mathbf{x})|, |\tau_0(\mathbf{x})|$ are bounded for all $\mathbf{x} \in \mathcal{X}$. Under standard regularity conditions, $\widehat{\xi} \xrightarrow{p} \xi_0$, for comparable populations $\mathcal{CP}$.*

We provide the proof in Appendix G.1.

### 3.3.2 Violation of the Assumptions and Sensitivity Analysis

In the previous section, we clarified that we can consistently estimate external robustness for populations that satisfy the common generalization assumptions. In this section, we show that, even when those common generalization assumptions are violated, the proposed estimator is a consistent estimator for the upper bound of the true external robustness. Due to space constraints, here we focus on the violation of Assumption 1 (Ignorability of Sampling and Treatment Effect Heterogeneity) and provide detailed discussions of the violation of Assumption 2 (Overlap) in Appendix A.2.2.

**Violation of Ignorability of Sampling and Treatment Effect Heterogeneity.** To consider the violation of Assumption 1, suppose that the assumption holds with respect to $\widetilde{\mathcal{P}}$ only

after conditioning on observed covariates $\mathbf{X}_i$ and unobserved variables $\mathbf{U}_i$. Because the CATE function with both observed and unobserved variables $\tau_0(\mathbf{X}_i, \mathbf{U}_i)$ has more variations than the one only with observed variables $\tau_0(\mathbf{X}_i)$, external robustness estimated with observed covariates $\mathbf{X}_i$ is over-estimated in general.

To formalize this point, we define a sensitivity parameter $\Gamma$ such that $-\Gamma \leq \tau_0(\mathbf{x}, \mathbf{u}) - \tau_0(\mathbf{x}) \leq \Gamma$ for all $\mathbf{x} \in \mathcal{X}$ and $\mathbf{u} \in \mathcal{U}$ where $\Gamma \geq 0$ and $\mathcal{U}$ is the support of the unobserved variables $\mathbf{U}$. When Assumption 1 holds with observed covariates $\mathbf{X}_i$, $\Gamma = 0$. When there is larger unobserved treatment effect heterogeneity, $\Gamma$ is larger. Using this formalization, we can write down the KL minimization problem in cases where Assumption 1 is violated.

$$\mathrm{KL}_0^\dagger(\Gamma) \;\coloneqq\; \min_{\tau_0(\mathbf{x},\mathbf{u}) \in \mathcal{F}(\Gamma)} \min_{\widetilde{\mathcal{P}} \in \mathcal{CP}^\dagger} \mathrm{KL}(\widetilde{\mathcal{P}} || \mathcal{P}_{\mathtt{exp}}) \qquad \text{s.t.} \quad \underbrace{\mathbb{E}\{\tau_0(\mathbf{X}_i, \mathbf{U}_i); \widetilde{\mathcal{P}}\}}_{\text{T-PATE constraint}} \leq 0$$

where $\mathcal{CP}^\dagger$ is a class of comparable populations defined with respect to $(\mathbf{X}_i, \mathbf{U}_i)$ and $\mathcal{F}(\Gamma)$ is a class of the CATE functions $\tau_0(\mathbf{x}, \mathbf{u})$ that satisfy $-\Gamma \leq \tau_0(\mathbf{x}, \mathbf{u}) - \tau_0(\mathbf{x}) \leq \Gamma$ for all $\mathbf{x} \in \mathcal{X}$ and $\mathbf{u} \in \mathcal{U}$. Therefore, when Assumption 1 is violated, the true external robustness is defined as $\xi_0^\dagger(\Gamma) \coloneqq 1 - \exp(-\mathrm{KL}_0^\dagger(\Gamma))$.

This setup generalizes the KL minimization problem in equation (8) in two ways. First, the T-PATE constraint is now averaged over the CATE $\tau_0(\mathbf{X}_i, \mathbf{U}_i)$ that includes unobserved variables $\mathbf{U}_i$. Second, because the CATE function including unobserved variables $\tau_0(\mathbf{X}_i, \mathbf{U}_i)$ is inherently unobserved, we consider the worst-case, i.e., minimizing the KL-divergence with respect to $\tau_0(\mathbf{x}, \mathbf{u})$ within the bound specified by a sensitivity parameter $\Gamma$.

Using the optimization theory, we show that $\xi_0^\dagger(\Gamma) \leq \xi_0$. Therefore, the proposed estimator is an estimator for the upper bound of the true external robustness. Importantly, if researchers are willing to specify $\Gamma$, they can explicitly conduct a sensitivity analysis; how much estimated external robustness changes depending on the value of $\Gamma$, which captures the degree of violation of Assumption 1. We provide the proof and describe details of the sensitivity analysis in Appendix A.2.1.

### 3.4 Extensions

We provide a number of extensions in Appendix B. We have so far assumed that researchers do not have any information about potential populations. This is a core advantage of our approach as researchers often do not have data on populations in many applications. However, in some scenarios, we have partial knowledge about population data, and we might want to estimate robustness only against "plausible" populations. For example, suppose education is an important moderator, and we know the "plausible" proportion of college graduates is between 30 and 50 % in relevant populations. In this case, we can estimate external robustness only

against populations that have a proportion of college graduates between 30 and 50 %. We discuss how to incorporate such partial knowledge about populations in Appendix B.1.

In Appendix B.2, we also discuss how to conduct external robustness analysis using subgroups. In Appendix B.3 and B.4, we examine how to extend our method to external validity in terms of contexts and to certain types of observational studies, respectively. Finally, while we provide asymptotic results in Section 3.3, we provide simulation results to investigate how our proposed methods work with finite samples in Appendix G.

# 4 Benchmarks and Interpretation

One of our main contributions is to provide a general framework to construct benchmarks for external robustness. As formally developed in Section 3, external robustness quantifies how much reweighting is required to explain away the T-PATE. To substantively interpret the degree of external robustness, a question is, "how much reweighting is large?" To answer this question, we rely on two types of surveys that political scientists have used and validated for many years; national surveys (e.g., ANES and CES) and samples from Amazon Mechanical Turk (MTurk).

The underlying idea is simple — to approximate a nationally representative population, reweighting required for national surveys is relatively small, while reweighting required for the MTurk samples is larger. Therefore, we use (a) the amount of reweighting required for national surveys to approximate nationally representative populations and (b) the amount of reweighting required for the MTurk samples to approximate nationally representative populations as benchmarks to substantively interpret the degree of external robustness.

In Section 4.4, we also discuss how to gain a more comprehensive picture of external robustness by investigating covariate profiles of the population for which the T-PATE is zero.

## 4.1 Data Sources

### 4.1.1 National Surveys

National surveys like the American National Election Studies (ANES) and the Cooperative Election Study (CES; formerly the Cooperative Congressional Election Study) spend significant efforts to obtain nationally representative samples. Yet, it is impossible to have perfectly representative samples due to non-response, attrition, and other problems. Therefore, to correct this deviation from an ideal probability sampling, national surveys offer survey weights.

The important point from our perspective is that reweighting required in national surveys is relatively small, in particular, compared to reweighting required for the MTurk samples we discuss below. Therefore, survey weights given in national surveys are useful benchmarks for "small" reweighting. This is the amount of reweighting required even when survey administra-

tors spend significant efforts on probability sampling.

We emphasize that it is possible that survey weights provided in national surveys are optimistic, and thus, the amount of reweighting required to approximate nationally representative populations might be underestimated. However, we will use this as a benchmark for "small" reweighting, so we view it as a conservative approach.

We rely on eight widely-used national surveys: American National Election Studies (ANES), Cooperative Election Study (CES), European Social Survey (ESS), Afrobarometer, Arab Barometer, Asian Barometer, Eurobarometer, and Latinobarometer. We investigate the five most recent waves in each survey. In total, we have 40 national surveys. We empirically show below that reweighting required for national surveys is small, as expected, and similar across the eight different surveys.

#### 4.1.2 Mechanical Turk Samples

The MTurk online samples have been widely used and validated in political science. Common findings are that the MTurk samples are often more representative of the U.S. population than in-person convenience samples but less representative than national surveys (Berinsky, Huber and Lenz, 2012). From our perspective, the MTurk samples are useful benchmarks to understand "moderate" reweighting. The MTurk samples are widely used, and thus, the amount of reweighting required for the MTurk samples to approximate a nationally representative population is common, while it is not as small as the one required for national surveys.

We consider four famous validation studies of the MTurk samples (Berinsky, Huber and Lenz, 2012; Huff and Tingley, 2015; Mullinix et al., 2015; Coppock, Leeper and Mullinix, 2018). Using covariates examined in each validation study, we compute weights required to approximate nationally representative populations. In total, we have 31 MTurk data sets. We show below that the amount of reweighting required for the MTurk samples is larger than that required for national surveys, and it is also similar across different validation studies.

### 4.2 Creating Benchmarks

We now formally discuss a framework for constructing benchmarks with different types of surveys. We want to emphasize that we are not estimating external robustness of surveys here. Rather, we use different levels of representativeness of well-known surveys to construct benchmarks for external robustness.

For each survey, we first construct survey weights required to approximate a nationally representative population. For national surveys, we rely on survey weights provided in the original surveys. As discussed above, these survey weights might be optimistic. However, we

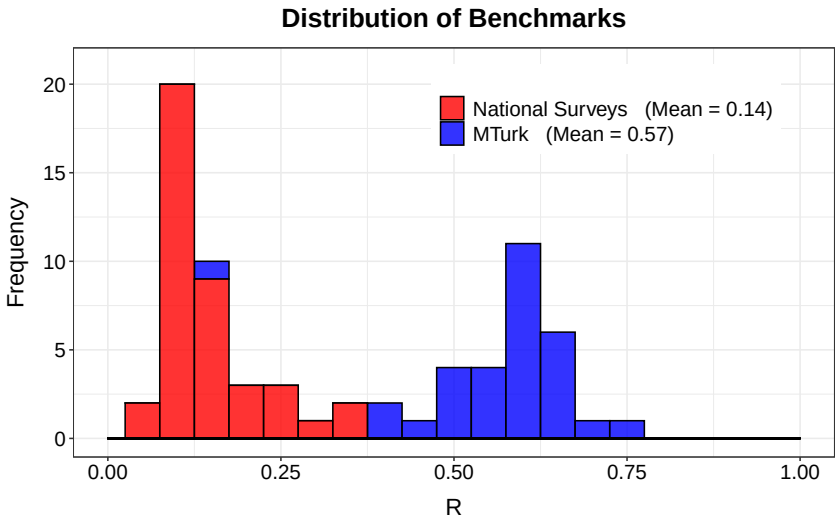



Figure 3: Benchmarks based on National Surveys and the MTurk Samples.

will use this as a benchmark for "small" reweighting, so we consider it to be a conservative choice. For the MTurk samples, we use entropy balancing (Deville and Särndal, 1992; Hainmueller, 2012) to estimate weights that match the marginal distributions of the MTurk samples and nationally representative samples. We consider the largest set of covariates common to both the nationally representative samples and the MTurk samples collected in the validation studies mentioned above. We provide details in Appendix C.

Given the weights in each survey, we quantify the amount of reweighting based on the same procedure used in Section 3. That is, we compute $R_s = 1 - \exp(-\text{KL}(\widehat{\mathbf{w}}_s))$, where $\text{KL}(\cdot)$ is the Kullback–Leibler divergence and $\widehat{\mathbf{w}}_s$ are the estimated survey weights in survey $s$.

Figure 3 reports the distributions of $R_s$ among national surveys and among the MTurk samples. Several points are worth noting. First, the amount of reweighting required for the national surveys is smaller than that required for the MTurk samples, as expected. The required amount of reweighting is also similar across the eight surveys. The mean is 0.14, and the median is 0.12. Second, the required amount of reweighting for the Mturk samples is larger and relatively similar across different validation studies. The mean is 0.57, and the median is 0.58.

In this paper, we recommend using 0.14 and 0.57 as default benchmarks for the "small" and "moderate" amount of reweighting. If researchers are worried that simply relying on these two values hides variations within the national surveys and the MTurk samples, they can also directly use the full distribution in Figure 3 to interpret the degree of estimated external robustness.

We emphasize that our proposed framework is general and is not tied to our choice of the national surveys and the MTurk samples. If necessary, researchers can use other relevant surveys to update benchmarks that are better tailored to their own applications. Researchers can also

update benchmarks over time to reflect the temporal change in the qualities of survey sampling.

### 4.3 Interpreting External Robustness against Benchmarks

How can we use these benchmarks to interpret external robustness substantively? If estimated external robustness is smaller than 0.14, the experimental results are only robust to populations that are very similar to the experimental sample. This is because 0.14 is equal to the amount of reweighting required for national surveys, which is small. If experimental findings are robust only to reweighting similar to that required for national surveys, it suggests that experimental results have low external robustness because causal estimates will be equal to zero as long as a hypothetical population of interest is slightly different from the experimental sample.

In contrast, suppose estimated external robustness is larger than 0.57. In this case, the experimental results are robust to populations that are relatively different from the experimental sample. This is because 0.57 is equal to the amount of reweighting required for the MTurk samples to approximate nationally representative populations, which is relatively large. This suggests that experimental findings have relatively high external robustness because causal estimates will be equal to zero only when the experimental sample is as different from a hypothetical population as the MTurk samples are from the U.S. general population.

When estimated external robustness is between 0.14 and 0.57, external robustness is moderate. The experimental results are robust to populations that are very similar, while they are not to populations that are relatively different from the experimental sample (the difference as large as that between the MTurk sample and the U.S. general population).

### 4.4 Interpreting External Robustness using Covariates

To provide further intuitive interpretation of external robustness, we show that researchers can also investigate covariate profiles of the population that is closest to the experimental sample among populations for which the T-PATE is equal to zero.[3] Combining this with a single summary statistic of external robustness, researchers can gain a more comprehensive picture.

In particular, we can compute $\overline{\mathbf{X}}_w = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}_i \mathbf{X}_i$ to estimate the means of covariates in the population for which the T-PATE is equal to zero. Here $\widehat{w}_i$ are the estimated weights from equation (6). Researchers can compare them to the means of covariates in the experiment $\overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$. Researchers can explicitly examine their similarity by reporting the standardized difference $(\overline{\mathbf{X}} - \overline{\mathbf{X}}_w)/\text{sd}(\mathbf{X})$ where $\text{sd}(\mathbf{X})$ is the standard deviation of covariates in the

[3]We consider the population *closest* to the experimental sample because it directly quantifies the minimum amount of changes in the covariate distribution required to make the T-PATE equal to zero, as we introduced in Section 3.2.

| | Experimental Sample | Population with T-PATE = 0 | | | Standardized Difference | | |
|---|---|---|---|---|---|---|---|
| | | $\widehat{\xi} = 0.11$ | $\widehat{\xi} = 0.35$ | $\widehat{\xi} = 0.65$ | $\widehat{\xi} = 0.11$ | $\widehat{\xi} = 0.35$ | $\widehat{\xi} = 0.65$ |
| $X_1$ | 1.01 | 0.94 | 0.87 | 0.72 | 0.13 | 0.28 | 0.57 |
| $X_2$ | 1.00 | 0.97 | 0.93 | 0.76 | 0.07 | 0.16 | 0.48 |
| $X_3$ | 0.99 | 0.95 | 0.90 | 0.68 | 0.07 | 0.17 | 0.61 |

Table 1: Examples of Covariate Profiles of Populations with the T-PATE = 0. *Note:* When external robustness is low, the T-PATE is zero even in populations similar to the experimental sample. When external robustness is high, the T-PATE is zero only in populations substantially different from the experimental sample.

experimental data.

If the population with the zero T-PATE has similar means of covariates to those of the experimental sample, this shows that the T-PATE is zero even in populations that are only slightly different from the experimental sample. In contrast, if the population with the zero T-PATE has distinct means of covariates from those of the experimental sample, it shows that causal conclusions are robust to a wide range of populations.

In Table 1, we illustrate this approach using the same simulation data from Figure 2. We report the means of three covariates $(X_1, X_2, X_3)$ in the experimental data and in populations that have the T-PATE equal to zero in addition to the standardized differences. We can see that when estimated external robustness is low ($\widehat{\xi} = 0.11$), the T-PATE is only robust to small changes in covariates (e.g., roughly 0.10 standard deviations in this example), that is, the T-PATE is zero even in populations similar to the experimental sample. When estimated external robustness is higher ($\widehat{\xi} = 0.65$), the T-PATE is robust to relatively large changes in covariates (e.g., roughly 0.50 standard deviations in this example), that is, the T-PATE is zero only in populations relatively different from the experimental sample.

We note that this investigation of covariates' means can only capture linear relationships between $\mathbf{X}$ and treatment effect heterogeneity, while the proposed measure of external robustness itself can capture complex non-linear relationships because we do not make any parametric assumption about the CATE function. Therefore, it is always important to report estimated external robustness (compared against benchmarks) in addition to the covariate profiles.

# 5 Practical Considerations

In this section, we summarize various practical considerations: when to use, how to use, common concerns about generalization assumptions, and potential limitations of the proposed method.

### 5.1 When to Use

The proposed approach is useful to estimate external robustness and characterize robustness to external validity bias whenever researchers use some forms of convenience samples or non-probability samples. This is typically the case in many field and lab experiments, and in many recent survey experiments using Mechanical Turk, Lucid, and other online platforms.

We emphasize that, even when researchers use national surveys or have experimental samples that are approximately similar to a certain target population, it is still useful to estimate external robustness for two reasons. First, as we clarified in the introduction, there are potentially many relevant target populations, and it is possible that readers might be interested in populations other than the target population analysts specified. Second, even if a target population is similar to the experimental samples, when external robustness is low (e.g., below 0.14, following the benchmark in the previous section), the T-PATE estimate can still be sensitive to very small changes in populations.

Unlike most existing methods for generalization that require some population data, researchers can estimate external robustness only with experimental data. Therefore, for any type of experiment, we recommend computing external robustness whenever there are concerns about external validity bias.

### 5.2 How to Use

To estimate external robustness, we require three simple steps.

- Step 1: Users specify covariates $\mathbf{X}_i$ that are important for treatment effect heterogeneity. Importantly, researchers can use all covariates measured in the experimental data.

- Step 2: Users estimate conditional average treatment effects (CATEs) using machine learning based estimators, such as causal forest, and then estimate external robustness as a measure between 0 and 1 (see equations (6) and (7)). We recommend reporting external robustness based on a point estimate as well as the one that incorporates uncertainty approximated by bootstrap (see Section 3.2).

- Step 3: Researchers can use benchmarks (e.g., 0.14 and 0.57) to substantively interpret the magnitude of the estimated external robustness (see Section 4.3). To provide further understanding of external robustness, researchers can also report covariate profiles of the population for which the T-PATE is equal to zero (see Section 4.4).

All the steps can be implemented via a single function call in the companion R package `exr`.

### 5.3 Common Concern: Generalization Assumptions

The most common concern is about the potential violation of the generalization assumptions (Assumptions 1 and 2). Several points are worth noting here. First, the ignorability of sampling and treatment effect heterogeneity assumption is more plausible in our proposed approach because researchers can use all covariates in the experiment than in traditional generalization methods where researchers are forced to use covariates measured both in the experimental and population data.

Second, even when the generalization assumptions are violated, we show that estimated external robustness has a simple interpretation. In particular, the proposed estimator is consistent to the upper bound of the true external robustness (see Section 3.3.2). Thus, regardless of whether those assumptions hold, low estimated external robustness reveals a lack of robustness to external validity bias. While high estimated external robustness does not guarantee high external validity, it is necessary for external validity. If researchers want to further improve external robustness analysis, they can also use an explicit sensitivity analysis to investigate how close the upper bound is to the true external robustness (see Section 3.3.2).

We emphasize that the above discussion does not mean that researchers can ignore the generalization assumptions. Rather, it is critical to think about the generalization assumptions carefully because the estimated upper bound of external robustness, while it is valid, can be too loose if their violation is severe. In practice, it is essential to measure moderators carefully in the experiment at the design stage. Measuring moderators is not only useful for understanding external robustness but also for increasing efficiency of estimating the SATE for internal validity.

### 5.4 Potential Limitations

For any method, it is essential to understand potential limitations. There are at least two particular instances where estimation of external robustness may be uninformative. The first scenario is where researchers only have an extremely limited number of covariates (e.g., one or two) in the experimental data. In this case, it is likely that many important moderators are left unadjusted. The estimated external robustness, while providing a valid upper bound, is therefore likely to be too large. To avoid this issue, we recommend collecting important moderators in the experiment.

The second scenario is when researchers want to estimate the exact magnitude of the T-PATE rather than assessing the sign of the T-PATE (e.g., whether the T-PATE is positive). This type of question is essential when researchers or policy makers want to conduct cost-benefit analyses to scale up the intervention in the real world because the cost-benefit calculation will be affected by the exact magnitude of the T-PATE. As described in Section 3, researchers can

choose any threshold to estimate external robustness such that the T-PATE is less than or equal to some substantively significant effect size. However, our framework of external robustness cannot point-identify the T-PATE itself. Therefore, when researchers are interested in making policy recommendations with cost-benefit analyses, we recommend collecting population data of interest and estimating the T-PATE directly (see effect-generalization in Egami and Hartman (2022)).

# 6 Empirical Application

In this section, we illustrate how to report external robustness using a large-scale field experiment by Domurat, Menashe and Yin (2021). We find that, while the SATE is precisely estimated to be positive, external robustness is relatively low. This highlights the importance of considering external validity together with internal validity. In Appendix D, we also offer another empirical application based on a survey experiment, and we find high external robustness (as high as 1), showing that there is a wide variation of external robustness in practice.

## 6.1 Background: Domurat, Menashe and Yin (2021)

At any one time, about 10 million people have health coverage through a marketplace created by the Patient Protection and Affordable Care Act. However, as a fraction of eligible households, the take-up rate of health insurance in those health benefits marketplaces is surprisingly low. Domurat, Menashe and Yin (2021) examine how reducing behavioral frictions impacts enrollment decisions. The original authors randomly assigned California households to receive letter interventions, designed to lower informational and psychological frictions that could hinder take-up in the state's health benefits marketplace, called Covered California.

The experimental sample is composed of households that were determined to be eligible for Covered California 2016 coverage, but had not selected any insurance plan. These households were self-selected to directly apply for Covered California or had applied for the state Medicaid program before.[4] In these situations, the most common concern of external validity is that experimental subjects who participated in the study might differ from the general population of uninsured people in California because many uninsured people have never applied for Covered California or the state Medicaid program. In addition, there are potentially many relevant target populations because a similar intervention can be useful for other populations. Given the huge policy implications of this field experiment, it is critical to think about whether and how experimental results in this study can be generalized to other populations. To answer this question, we estimate external robustness to investigate how robust experimental results in this

[4]Please see the original paper as well as Appendix E.1 for more details about sampling.

study are to external validity bias.

As usually done in field experiments, households were randomly assigned to treatment and control groups. Households in the treatment group received a reminder letter about health insurance. These letters were designed to reduce behavioral frictions that potentially lower the take-up of insurance. For example, the most basic version showed the open enrollment deadline, general benefits of insurance, and the Covered California website and telephone number where they could shop for plans. While the original authors implemented four different versions of letters, they found similar causal effects of these interventions, and thus, we do not distinguish different versions below, as in the main analysis of the original paper. Households in the control group received no direct communication beyond the generic outreach and state-wide marketing activities used by Covered California, which represented a status-quo.

The original authors measured a binary variable indicating whether households took up insurance before the open enrollment deadline as the main outcome variable. To estimate the SATE of the reminder letter, the original authors used the difference-in-mean as well as a linear regression that includes the following observed covariates to improve efficiency without introducing bias: age, race, household income, marital status, family size, number of kids, language preferences, and the age-based community-rating premium ratio. Due to this study's large sample size of 87,394, the SATE is precisely estimated to be positive. With the difference-in-means, the SATE estimate is 1.3 percentage points (95% CI = $[0.8, 1.7]$; p-value = 0.00), and with a covariate-adjusted regression estimator, the SATE estimate is 1.2 percentage points (95% CI = $[0.8, 1.7]$; p-value = 0.00). Within the experiment, the original authors found clear evidence of a positive causal effect on average.

### 6.2 External Robustness

To estimate external robustness, we rely on the aforementioned eight observed covariates that the original authors used. Substantively, given that we adjust for a host of moderators relevant for treatment effect heterogeneity such as age, race, household income, and family size, it is likely that Assumptions 1 and 2 hold for many relevant populations. However, it is also reasonable to be concerned about the potential violation of these generalization assumptions. We discuss below how to interpret results when the generalization assumptions might be violated.

Using the variables selected above, we estimate external robustness using the experimental data alone. To estimate the CATEs as the intermediate step, we rely on one of the most popular CATE estimators, causal forest (Wager and Athey, 2018). We note that researchers can use any other CATE estimators as well. All steps were implemented via a single function call in the companion R package `exr`.

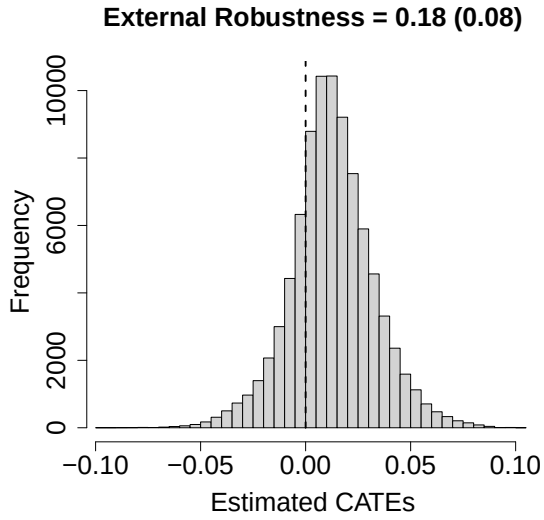



Figure 4: External Robustness and Distribution of Estimated CATEs for Domurat, Menashe and Yin (2021). *Note:* The estimated external robustness is 0.18. The estimated external robustness that incorporates uncertainty is 0.08, reported within parentheses.

Figure 4 shows the estimated external robustness and the distribution of estimated CATEs. The estimated external robustness is 0.18, which is only slightly larger than the benchmark for "small" reweighting (0.14) and is much smaller than the benchmark for "moderate" reweighting (0.57). When we take into account uncertainty (i.e., external robustness of the lower limit of the 95% confidence interval of the T-PATE, which is approximated by bootstrap), we find the estimated external robustness to be 0.08, which is smaller than the benchmark for "small" reweighting (0.14). This implies the T-PATE estimate is only robust to populations that are very similar to the experimental sample. As visually clear from Figure 4, this is because there is a large amount of treatment effect heterogeneity that ranges from negative to positive values.

Importantly, if users are concerned about potential violation of the generalization assumptions, they can interpret the estimated external robustness to be the upper bound, implying that the true external robustness is even lower. Overall, this estimation of external robustness shows that the T-PATE estimate is not robust to external validity bias in general.

Finally, to substantively understand how far the experimental results can be generalized, we also report the means of covariates in the closest population for which the T-PATE is equal to zero. In Table 2, the first column shows the means of key covariates in the experimental data, the second represents those in the closest population for which the T-PATE is equal to zero, and the third column shows the standardized difference. As we found with the estimated external robustness, we can see that the T-PATE is no longer positive even in populations that are only slightly different from the experimental sample. For most covariates, means are essentially the

| | Experimental Sample | Population with T-PATE = 0 | Standardized Difference |
|---|---|---|---|
| Mean age within household | 37.65 | 39.59 | -0.13 |
| Household size | 2.16 | 2.18 | -0.01 |
| White | 0.26 | 0.26 | -0.02 |
| Black | 0.05 | 0.04 | 0.00 |
| Latino | 0.43 | 0.43 | -0.00 |
| Asian | 0.12 | 0.13 | -0.01 |
| Log(Income) | 5.50 | 5.50 | -0.00 |

Table 2: Means of Key Covariates in the Experimental Sample and the Closest Population whose T-PATE is equal to zero. *Note:* Income is reported as percent of federal poverty limit, following the original paper.

same as those of the experimental sample, and even for age, we are only robust to the change in 0.13 of the standard deviation.

### 6.3 Additional Empirical Application and Validation

We provide additional empirical results in the Appendix to provide a more comprehensive picture of how external robustness works in practice.

In Appendix D, we provide another empirical application based on a survey experiment by Johnston and Ballard (2016). In contrast to the application we analyzed above, we find high external robustness (as high as 1), showing that there is a wide variation of external robustness in practice. Thus, it is important to explicitly estimate external robustness in each application instead of assuming high or low external robustness a priori without empirical evidence.

In Appendix F, we provide an empirical validation study using 13 pairs of original and replication experiments collected by Coppock, Leeper and Mullinix (2018). We empirically find that our proposed measure of external robustness can accurately assess the external validity of experiments only using the experimental data without requiring population data.

## 7 Concluding Remarks

The external validity of randomized experiments is essential for accumulating knowledge in the social sciences. Most existing methods aim to estimate the T-PATE by approximating a particular target population. Despite their importance, it is sometimes difficult to implement such methods because analysts and skeptics might not agree on a particular choice of the population or because it is infeasible to obtain a rich set of covariates for the selected target population. For these practical reasons, few applications include formal analysis of external

validity.

To tackle this practical challenge, we propose a measure of external robustness by estimating how much different a population should be from the experimental sample to explain away the T-PATE. This quantifies the robustness of experimental results to external validity bias. Unlike existing methods for estimating the T-PATE, estimation of external robustness only requires experimental data and no population data. Thus, researchers can estimate external robustness for any experimental study. We prove that the proposed estimator is consistent to the true external robustness under common generalization assumptions and, even more importantly, is consistent to the upper bound even when those assumptions are violated. Finally, we provide simple benchmarks based on national surveys and MTurk samples to help interpret the degree of external robustness in any given application.

## References


Andrews, Isaiah and Emily Oster. 2019. "A Simple Approximation for Evaluating External Validity Bias." *Economics Letters* 178:58–62.

Bareinboim, Elias and Judea Pearl. 2016. "Causal Inference and the Data-Fusion Problem." *Proceedings of the National Academy of Sciences* 113(27):7345–7352.

Berinsky, Adam J, Gregory A Huber and Gabriel S Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3):351–368.

Blair, Graeme and Gwyneth McClendon. 2020. Experiments in Multiple Contexts. In *Handbook of Experimental Political Science*, ed. Donald P. Green and James Druckman. Cambridge University Press.

Cole, Stephen R and Elizabeth A Stuart. 2010. "Generalizing Evidence From Randomized Clinical Trials to Target PopulationsThe ACTG 320 Trial." *American Journal of Epidemiology* 172(1):107–115.

Cooper, Harris, Larry V Hedges and Jeffrey C Valentine. 2019. *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation.

Coppock, Alexander, Thomas J Leeper and Kevin J Mullinix. 2018. "Generalizability of Heterogeneous Treatment Effect Estimates Across Samples." *Proceedings of the National Academy of Sciences* 115(49):12441–12446.

Dahabreh, Issa J, Sarah E Robertson, Eric J Tchetgen Tchetgen, Elizabeth A Stuart and Miguel A Hernán. 2019. "Generalizing Causal Inferences From Individuals In Randomized Trials to All Trial-Eligible Individuals." *Biometrics* 75(2):685–694.

Deaton, Angus and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* .

Deville, Jean-Claude and Carl-Erik Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87(418):376–382.

Domurat, Richard, Isaac Menashe and Wesley Yin. 2021. "The Role of Behavioral Frictions in Health Insurance Marketplace Enrollment and Risk: Evidence From A Field Experiment." *American Economic Review* 111(5):1549–74.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D Hyde et al. 2019. "Voter Information Campaigns and Political Accountability: Cumulative Findings from a Preregistered Meta-Analysis of Coordinated Trials." *Science Advances* 5(7):eaaw2612.

Egami, Naoki and Erin Hartman. 2021. "Covariate Selection for Generalizing Experimental Results: Application to a Large-Scale Development Program in Uganda." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 184(4):1524–1548.

Egami, Naoki and Erin Hartman. 2022. "Elements of External Validity: Framework, Design, and Analysis." *American Political Science Review* .

Findley, Michael G, Kyosuke Kikuta and Michael Denly. 2020. "External Validity." *Annual Review of Political Science* .

Gerber, Alan S and Donald P Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. WW Norton.

Green, Donald P and Holger L Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3):491–511.

Grimmer, Justin, Solomon Messing and Sean J Westwood. 2017. "Estimating Heterogeneous Treatment Effects and The Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25(4):413–434.

Gupta, Suyash and Dominik Rothenhäusler. 2021. "The $s$-value: Evaluating Stability with respect to Distributional Shifts." *arXiv preprint arXiv:2105.03067* .

Hainmueller, Jens. 2012. "Entropy Balancing For Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20(1):25–46.

Hartman, Erin, Richard Grieve, Roland Ramsahai and Jasjeet S Sekhon. 2015. "From Sample Average Treatment Effect to Population Average Treatment Effect on the Treated." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 178(3):757–778.

Hill, Jennifer L. 2012. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1):217–240.

Huang, Melody. 2022. "Sensitivity Analysis in the Generalization of Experimental Results." *arXiv preprint arXiv:2202.03408* .

Huff, Connor and Dustin Tingley. 2015. ""Who Are These People?" Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents." *Research & Politics* 2(3):2053168015604648.

Imai, Kosuke, Gary King and Elizabeth A Stuart. 2008. "Misunderstandings Between Experimentalists and Observationalists About Causal Inference." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(2):481–502.

Imai, Kosuke and Marc Ratkovic. 2013. "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation." *Annals of Applied Statistics* 7(1):443–470.

Johnston, Christopher D and Andrew O Ballard. 2016. "Economists and Public Opinion: Expert Consensus and Economic Policy Judgments." *The Journal of Politics* 78(2):443–456.

Kennedy, Edward H. 2020. "Optimal Doubly Robust Estimation of Heterogeneous Causal Effects." *arXiv preprint arXiv:2004.14497* .

Kern, Holger L, Elizabeth A Stuart, Jennifer Hill and Donald P Green. 2016. "Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations." *Journal of Research on Educational Effectiveness* 9(1):103–127.

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel and Bin Yu. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects using Machine Learning." *Proceedings of the national academy of sciences* 116(10):4156–4165.

Miratrix, Luke W, Jasjeet S Sekhon, Alexander G Theodoridis and Luis F Campos. 2018. "Worth Weighting? How to Think About and Use Weights in Survey Experiments." *Political Analysis* 26(3):275–291.

Mullinix, Kevin J, Thomas J Leeper, James Druckman and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2(2):109–138.

Munger, Kevin. 2019. "Knowledge Decays: Temporal Validity and SocialScience in a Changing World." *Working Paper* .

Nguyen, Trang Quynh, Cyrus Ebnesajjad, Stephen R Cole and Elizabeth A Stuart. 2017. "Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects." *The Annals of Applied Statistics* 11(1):225–247.

Nie, Xinkun, Guido Imbens and Stefan Wager. 2021. "Covariate Balancing Sensitivity Analysis for Extrapolating Randomized Trials across Locations." *arXiv preprint arXiv:2112.04723* .

Shadish, William R, Thomas D Cook and Donald T Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Slough, Tara and Scott A Tyson. 2022. "External Validity and Meta-Analysis." *American Journal of Political Science* .

Spini, Pietro Emilio. 2021. "Robustness, Heterogeneous Treatment Effects and Covariate Shifts." *arXiv preprint arXiv:2112.09259* .

Stuart, Elizabeth A, Stephen R Cole, Catherine P Bradshaw and Philip J Leaf. 2011. "The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174(2):369–386.

Tipton, Elizabeth. 2013. "Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts." *Journal of Educational and Behavioral Statistics* 38(3):239–266.

Tipton, Elizabeth. 2014. "How Generalizable is Your Experiment? An Index for Comparing Experimental Samples and Populations." *Journal of Educational and Behavioral Statistics* 39(6):478–501.

Wager, Stefan and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113(523):1228–1242.

Wilke, Anna and Macartan Humphreys. 2020. Field Experiments, Theory, and External Validity. In *The SAGE Handbook of Research Methods in Political Science and International Relations*, ed. Luigi Curini and Robert Franzese. Transaction Publishers.

# Online Supplementary Appendix

## Quantifying Robustness to External Validity Bias

## Table of Contents

# A Proofs and Methodological Details

## A.1 Proof of Consistency: Theorem 1

### A.1.1 Setup

As in Section 2.2, we define

$$\tau_0(\mathbf{X}_i) \;\coloneqq\; \mathbb{E}\{Y_i(1) - Y_i(0) \mid \mathbf{X}_i; \mathcal{P}_{\texttt{exp}}\}. \tag{1}$$

For comparable populations, $\mathbb{E}\{Y_i(1) - Y_i(0) \mid \mathbf{X}_i; \mathcal{P}_{\texttt{exp}}\} = \mathbb{E}\{Y_i(1) - Y_i(0) \mid \mathbf{X}_i; \widetilde{\mathcal{P}}\}$. $\widehat{\tau}(\mathbf{X}_i)$ is an estimator for $\tau_0(\mathbf{X}_i)$.

We consider the following minimization problem.

$$\text{KL}_0 \coloneqq \min_{\widetilde{\mathcal{P}} \in \mathcal{CP}} \text{KL}(\widetilde{\mathcal{P}} || \mathcal{P}_{\texttt{exp}}) \qquad \text{s.t.} \quad \mathbb{E}\{\tau_0(\mathbf{X}_i); \widetilde{\mathcal{P}}\} \leq 0 \tag{2}$$

where

$$\text{KL}(\widetilde{\mathcal{P}} || \mathcal{P}_{\texttt{exp}}) \coloneqq \int \log\left(\frac{d\widetilde{\mathcal{P}}}{d\mathcal{P}_{\texttt{exp}}}\right) d\widetilde{\mathcal{P}}.$$

The true external robustness is defined to be $\xi_0 \coloneqq 1 - \exp(-\text{KL}_0)$.

Using the dual of equation (2) (Boyd and Vandenberghe, 2004), we can rewrite external robustness as

$$\begin{aligned} \xi_0 &= 1 - \min_{\lambda} \mathbb{E}\{\exp(-\lambda \tau_0(\mathbf{X}_i)); \mathcal{P}_{\texttt{exp}}\} \quad \text{s.t.} \quad \lambda \geq 0. \\ &= 1 - \mathbb{E}\{\exp(-\lambda_0 \tau_0(\mathbf{X}_i)); \mathcal{P}_{\texttt{exp}}\} \end{aligned}$$

where

$$\lambda_0 \;=\; \operatorname*{argmin}_{\lambda \in \Lambda} \mathbb{E}\{\exp(-\lambda \tau_0(\mathbf{X}_i))\},$$

and $\Lambda$ guarantees that $\lambda \geq 0$.

The empirical version of the primary problem is written as follows.

$$\begin{aligned} \widehat{\text{KL}} \;\coloneqq\; & \min_{\mathbf{w}} \; \frac{1}{n} \sum_{i=1}^{n} w_i \log(w_i) \\ & \text{s.t.} \quad \frac{1}{n} \sum_{i=1}^{n} w_i \widehat{\tau}(\mathbf{X}_i) \leq 0 \quad \sum_{i=1}^{n} w_i = n, \quad w_i \geq 0. \end{aligned} \tag{3}$$

Using the dual of equation (3), we have

$$\begin{aligned} \exp(-\widehat{\text{KL}}) \;\coloneqq\; & \min_{\lambda} \frac{1}{n} \sum_{i=1}^{n} \exp(-\lambda \widehat{\tau}(\mathbf{X}_i)) \\ & \text{s.t.} \quad \lambda \geq 0. \end{aligned} \tag{4}$$

Therefore, we have

$$\widehat{\xi} \;=\; 1 - \frac{1}{n}\sum_{i=1}^{n} \exp(-\widehat{\lambda}\widehat{\tau}(\mathbf{X}_i)) \tag{5}$$

where

$$\widehat{\lambda} \;=\; \underset{\lambda\in\Lambda}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \exp(-\lambda\widehat{\tau}(\mathbf{X}_i)). \tag{6}$$

Therefore, we need to prove

$$\frac{1}{n}\sum_{i=1}^{n} \exp(-\widehat{\lambda}\widehat{\tau}(\mathbf{X}_i)) \xrightarrow{p} \mathbb{E}\{\exp(-\lambda_0\tau_0(\mathbf{X}_i)); \mathcal{P}_{\texttt{exp}}\}. \tag{7}$$

**Regularity Conditions.** We assume the following standard regularity conditions. (i) $\mathcal{F}$ is Glivenko-Cantelli where $\mathcal{F}$ is a class of functions used to estimate $\tau_0(\mathbf{X}_i)$. (ii) $\Lambda$ is compact. (iii) $\exp(-\lambda\widehat{\tau}(\mathbf{X}_i))$ is continuous at $\lambda \in \Lambda$ with probability one. (iv) $\exp(-\lambda\widehat{\tau}(\mathbf{X}_i))$ is dominated by a function $G(\mathbf{X}_i)$, and $\mathbb{E}\{G(\mathbf{X}_i)\}$ is bounded. (v) $|\widehat{\tau}(\mathbf{X}_i)|$ and $|\tau_0(\mathbf{X}_i)|$ are bounded.

We also assume that the conditional ATE estimator $\widehat{\tau}$ is consistent in $L_2$, i.e., $\mathbb{E}\{(\widehat{\tau}(\mathbf{X}_i) - \tau_0(\mathbf{X}_i))^2\} \xrightarrow{p} 0$, which is satisfied for most well-known machine learning based estimators.

**Proof Structure.** To prove equation (7), we show the following three steps.

1. Prove $\sup_{\lambda\in\Lambda} \left|\frac{1}{n}\sum_{i=1}^{n} \exp(-\lambda\widehat{\tau}(\mathbf{X}_i)) - \mathbb{E}\{\exp(-\lambda\tau_0(\mathbf{X}_i)); \mathcal{P}_{\texttt{exp}}\}\right| \xrightarrow{p} 0$.
2. Prove $\widehat{\lambda} \xrightarrow{p} \lambda_0$.
3. Prove $\frac{1}{n}\sum_{i=1}^{n} \exp(-\widehat{\lambda}\widehat{\tau}(\mathbf{X}_i)) \xrightarrow{p} \mathbb{E}\{\exp(-\lambda_0\tau_0(\mathbf{X}_i)); \mathcal{P}_{\texttt{exp}}\}$.

Expectations we use are over $\mathcal{P}_{\texttt{exp}}$ unless otherwise noted, so for notational simplicity, we omit $\mathcal{P}_{\texttt{exp}}$ below.

### A.1.2 Proof

**Step 1**

**Step 1.1** Assume that $\mathcal{F}$ is Glivenko-Cantelli where $\mathcal{F}$ is a class of functions used to estimate $\tau_0(\mathbf{X}_i)$. Most machine learning based CATE estimators are within a Glivenko-Cantelli class, while not in a Donsker class. Therefore, for a given $\lambda \in \Lambda$,

$$\sup_{\tau\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \exp(-\lambda\tau(\mathbf{X}_i)) - \mathbb{E}\{\exp(-\lambda\tau(\mathbf{X}_i))\}\right| \xrightarrow{p} 0.$$

Thus, for a given $\lambda \in \Lambda$,

$$\frac{1}{n}\sum_{i=1}^{n} \exp(-\lambda\widehat{\tau}(\mathbf{X}_i)) - \mathbb{E}\{\exp(-\lambda\widehat{\tau}(\mathbf{X}_i))\} \xrightarrow{p} 0.$$

Because (a) $\Lambda$ is compact, (b) $\exp(-\lambda\widehat{\tau}(\mathbf{X}_i))$ is continuous at $\lambda \in \Lambda$ with probability one, and (c) $\exp(-\lambda\widehat{\tau}(\mathbf{X}_i))$ is dominated by a function $G(\mathbf{X}_i)$, and (d) $\mathbb{E}\{G(\mathbf{X}_i)\}$ is bounded, we have the uniform convergence.

$$\sup_{\lambda\in\Lambda}\left|\frac{1}{n}\sum_{i=1}^{n}\exp(-\lambda\widehat{\tau}(\mathbf{X}_i)) - \mathbb{E}\{\exp(-\lambda\widehat{\tau}(\mathbf{X}_i))\}\right| \xrightarrow{p} 0.$$

**Step 1.2** Due to the Minkowski inequality,

$$\begin{aligned}
&\sup_{\lambda\in\Lambda}\left|\frac{1}{n}\sum_{i=1}^{n}\exp(-\lambda\widehat{\tau}(\mathbf{X}_i)) - \mathbb{E}\{\exp(-\lambda\tau_0(\mathbf{X}_i))\}\right| \\
\leq\ &\sup_{\lambda\in\Lambda}\left|\frac{1}{n}\sum_{i=1}^{n}\exp(-\lambda\widehat{\tau}(\mathbf{X}_i)) - \mathbb{E}\{\exp(-\lambda\widehat{\tau}(\mathbf{X}_i))\}\right| + \sup_{\lambda\in\Lambda}\left|\mathbb{E}\{\exp(-\lambda\widehat{\tau}(\mathbf{X}_i))\} - \mathbb{E}\{\exp(-\lambda\tau_0(\mathbf{X}_i))\}\right|
\end{aligned}$$

Because Step 1.1 shows that the first term is $o_p(1)$, the main goal here is to show that

$$\sup_{\lambda\in\Lambda}\left|\mathbb{E}\{\exp(-\lambda\widehat{\tau}(\mathbf{X}_i))\} - \mathbb{E}\{\exp(-\lambda\tau_0(\mathbf{X}_i))\}\right| \xrightarrow{p} 0.$$

We analyze the term inside, $\mathbb{E}\{\exp(-\lambda\widehat{\tau}(\mathbf{X}_i))\} - \mathbb{E}\{\exp(-\lambda\tau_0(\mathbf{X}_i))\}$. Because $|\widehat{\tau}(\mathbf{X}_i)|$ and $|\tau_0(\mathbf{X}_i)|$ are bounded,

$$\begin{aligned}
&\left|\mathbb{E}\{\exp(-\lambda\widehat{\tau}(\mathbf{X}_i))\} - \mathbb{E}\{\exp(-\lambda\tau_0(\mathbf{X}_i))\}\right| \\
\leq\ &\mathbb{E}\left\{\exp(-\lambda\tau_0(\mathbf{X}_i)) \times |\exp(-\lambda(\widehat{\tau}(\mathbf{X}_i) - \tau_0(\mathbf{X}_i))) - 1\ |\right\} \\
\leq\ &C_\lambda\mathbb{E}\left\{|\exp(-\lambda(\widehat{\tau}(\mathbf{X}_i) - \tau_0(\mathbf{X}_i))) - 1\ |\right\} \\
\leq\ &C_\lambda\mathbb{E}\left\{|\{1 + D_\lambda(\widehat{\tau}(\mathbf{X}_i) - \tau_0(\mathbf{X}_i))\} - 1\ |\right\} \\
\leq\ &C_\lambda D_\lambda\mathbb{E}\left\{|(\widehat{\tau}(\mathbf{X}_i) - \tau_0(\mathbf{X}_i)|\right\}
\end{aligned}$$

where the first inequality follows from $\exp(\lambda\tau_0(\mathbf{X}_i)) > 0$ and the second from bounded $|\tau_0(\mathbf{X}_i)|$ and we define $C_\lambda$ as a function of $\lambda$ such that $\exp(-\lambda\tau_0(\mathbf{X}_i)) \leq C_\lambda$. Because $\widehat{\tau}(\mathbf{X}_i) - \tau_0(\mathbf{X}_i)$ is bounded, there exists $D_\lambda > 0$, which is a function of $\lambda$ that satisfies $|1+D_\lambda(\widehat{\tau}(\mathbf{X}_i)-\tau_0(\mathbf{X}_i))-1| \geq |\exp(-\lambda(\widehat{\tau}(\mathbf{X}_i)-\tau_0(\mathbf{X}_i)))-1|$, which leads to the third inequality. The fourth inequality follows from $D_\lambda > 0$.

Therefore, given that $\Lambda$ is compact, when $\mathbb{E}\left\{|(\widehat{\tau}(\mathbf{X}_i) - \tau_0(\mathbf{X}_i)|^2\right\} \xrightarrow{p} 0$,

$$\begin{aligned}
&\sup_{\lambda\in\Lambda}\left|\mathbb{E}\{\exp(-\lambda\widehat{\tau}(\mathbf{X}_i))\} - \mathbb{E}\{\exp(-\lambda\tau_0(\mathbf{X}_i))\}\right| \\
\leq\ &\sup_{\lambda\in\Lambda} C_\lambda D_\lambda\mathbb{E}\left\{|(\widehat{\tau}(\mathbf{X}_i) - \tau_0(\mathbf{X}_i)|\right\} \\
=\ &\widetilde{C}\widetilde{D}\mathbb{E}\left\{|(\widehat{\tau}(\mathbf{X}_i) - \tau_0(\mathbf{X}_i)|\right\} \\
=\ &o_p(1)
\end{aligned}$$

where $\widetilde{C}$ and $\widetilde{D}$ are the supremum of $C_\lambda$ and $D_\lambda$.

**Step 2**

Define $Q_n(\lambda) \coloneqq \frac{1}{n}\sum_{i=1}^n \exp(-\lambda\widehat{\tau}(\mathbf{X}_i))$ and $Q_0(\lambda) \coloneqq \mathbb{E}\{\exp(-\lambda\tau_0(\mathbf{X}_i))\}$. Then, by definition,

$$\begin{aligned}\lambda_0 &= \operatorname*{argmin}_{\lambda\in\Lambda} Q_0(\lambda) \\ \widehat{\lambda} &= \operatorname*{argmin}_{\lambda\in\Lambda} Q_n(\lambda)\end{aligned}$$

and, $\lambda_0$ is the unique minimizer of $Q_0(\lambda)$. From Step 1, we have

$$\sup_{\lambda\in\Lambda}\left|Q_n(\lambda) - Q_0(\lambda)\right| \xrightarrow{p} 0. \tag{8}$$

Therefore, we have (i) $Q_n(\lambda)$ is uniquely minimized at $\lambda_0$, (ii) parameter space $\Lambda$ is compact, (iii) $Q_n(\lambda)$ is continuous, and (iv) the uniform convergence (equation (8)). Thus, using Theorem 2.1 in Newey and McFadden (1994), $\widehat{\lambda} \xrightarrow{p} \lambda_0$.

**Step 3**

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \exp(-\widehat{\lambda}\widehat{\tau}(\mathbf{X}_i)) - \mathbb{E}\{\exp(-\lambda_0\tau_0(\mathbf{X}_i))\} \\ =\ &\left\{\frac{1}{n}\sum_{i=1}^n \exp(-\widehat{\lambda}\widehat{\tau}(\mathbf{X}_i)) - \mathbb{E}\{\exp(-\widehat{\lambda}\tau_0(\mathbf{X}_i))\}\right\} + \left\{\mathbb{E}\{\exp(-\widehat{\lambda}\tau_0(\mathbf{X}_i))\} - \mathbb{E}\{\exp(-\lambda_0\tau_0(\mathbf{X}_i))\}\right\}\end{aligned}$$

The first term is $o_p(1)$ because $\sup_{\lambda\in\Lambda}\left|\frac{1}{n}\sum_{i=1}^n \exp(-\lambda\widehat{\tau}(\mathbf{X}_i)) - \mathbb{E}\{\exp(-\lambda\tau_0(\mathbf{X}_i))\}\right| \xrightarrow{p} 0$ from Step 1. The second term is $o_p(1)$ due to continuous mapping theorem and $\widehat{\lambda} \xrightarrow{p} \lambda_0$ from Step 2. □

## A.2 Violation of the Generalization Assumptions

### A.2.1 Ignorability of Sampling and Treatment Effect Heterogeneity

**Optimization Problem**

We first formally define a class of populations $\mathcal{CP}^\dagger$ to be distributions $\widetilde{\mathcal{P}}$ that satisfy the following two conditions for a certain subset of $(\mathbf{X}, \mathbf{U})$, which we denote with $(\mathbf{X}_i^\dagger, \mathbf{U}_i^\dagger)$.

$$\mathbb{E}(Y_i(1) - Y_i(0) \mid \mathbf{X}_i^\dagger = \mathbf{x}^\dagger, \mathbf{U}_i^\dagger = \mathbf{u}^\dagger; \mathcal{P}_{\texttt{exp}}) = \mathbb{E}(Y_i(1) - Y_i(0) \mid \mathbf{X}_i^\dagger = \mathbf{x}^\dagger, \mathbf{U}_i^\dagger = \mathbf{u}^\dagger; \widetilde{\mathcal{P}}), \tag{9}$$

$$\Pr(\mathbf{X}_i^\dagger = \mathbf{x}^\dagger, \mathbf{U}_i^\dagger = \mathbf{u}^\dagger; \widetilde{\mathcal{P}}) > 0 \Longrightarrow \Pr(\mathbf{X}_i^\dagger = \mathbf{x}^\dagger, \mathbf{U}_i^\dagger = \mathbf{u}^\dagger; \mathcal{P}_{\texttt{exp}}) > 0. \tag{10}$$

When the ignorability of sampling and treatment effect heterogeneity assumption is violated, we consider the following general KL minimization problem.

$$\begin{aligned}&\mathrm{KL}_0^\dagger(\Gamma) \\ \coloneqq\ &\min_{\tau_0(\mathbf{x},\mathbf{u})\in\mathcal{F}(\Gamma)} \min_{\widetilde{\mathcal{P}}\in\mathcal{CP}^\dagger} \mathrm{KL}(\widetilde{\mathcal{P}}||\mathcal{P}_{\texttt{exp}}) \qquad \text{s.t.} \quad \mathbb{E}\{\tau_0(\mathbf{X}_i, \mathbf{U}_i); \widetilde{\mathcal{P}}\} \leq 0\end{aligned} \tag{11}$$

$$= \min_{\tau_0(\mathbf{x},\mathbf{u})\in\mathcal{F}(\Gamma)} \min_{\widetilde{\mathcal{P}}\in\mathcal{CP}^\dagger} \mathrm{KL}(\widetilde{\mathcal{P}}||\mathcal{P}_{\mathtt{exp}}) \qquad \text{s.t.} \quad \mathbb{E}\{\tau_0(\mathbf{X}_i);\widetilde{\mathcal{P}}\} \leq -\mathbb{E}\{\tau_0(\mathbf{X}_i,\mathbf{U}_i) - \tau_0(\mathbf{X}_i);\widetilde{\mathcal{P}}\},$$

where $\mathcal{F}(\Gamma)$ is a class of the conditional ATE function $\tau_0(\mathbf{x},\mathbf{u})$ that satisfies $-\Gamma \leq \tau_0(\mathbf{x},\mathbf{u}) - \tau_0(\mathbf{x}) \leq \Gamma$ for all $\mathbf{x}\in\mathcal{X}$ and $\mathbf{u}\in\mathcal{U}$.

Given that the objective function is a decreasing function of $-\mathbb{E}\{\tau_0(\mathbf{X}_i,\mathbf{U}_i) - \tau_0(\mathbf{X}_i);\widetilde{\mathcal{P}}\}$ and $-\mathbb{E}\{\tau_0(\mathbf{X}_i,\mathbf{U}_i) - \tau_0(\mathbf{X}_i);\widetilde{\mathcal{P}}\} \leq \Gamma$, we have

$$\mathrm{KL}_0^\dagger(\Gamma) \;=\; \min_{\widetilde{\mathcal{P}}\in\mathcal{CP}^\dagger} \mathrm{KL}(\widetilde{\mathcal{P}}||\mathcal{P}_{\mathtt{exp}}) \qquad \text{s.t.} \quad \mathbb{E}\{\tau_0(\mathbf{X}_i);\widetilde{\mathcal{P}}\} \leq \Gamma.$$

Because $\mathcal{CP}^\dagger$ is a superset of $\mathcal{CP}$ and the main component of the constraint $\mathbb{E}\{\tau_0(\mathbf{X}_i);\widetilde{\mathcal{P}}\}$ is a function of $\mathbf{X}_i$ alone,

$$\mathrm{KL}_0^\dagger(\Gamma) \;=\; \min_{\widetilde{\mathcal{P}}\in\mathcal{CP}} \mathrm{KL}(\widetilde{\mathcal{P}}||\mathcal{P}_{\mathtt{exp}}) \qquad \text{s.t.} \quad \mathbb{E}\{\tau_0(\mathbf{X}_i);\widetilde{\mathcal{P}}\} \leq \Gamma. \tag{12}$$

Because $\mathrm{KL}_0^\dagger(\Gamma) = \mathrm{KL}_0$ when $\Gamma = 0$ and $\mathrm{KL}_0^\dagger(\Gamma)$ is an increasing function of $\Gamma \geq 0$, this shows that the proposed estimator is an estimator for the upper bound of the true external robustness.

**Sensitivity Analysis**

To understand how external robustness changes depending on the degree of violation of the assumption, researchers can conduct an explicit sensitivity analysis. In particular, researchers can choose different values of $\Gamma$ to consider a wide range of scenarios where the ignorability of sampling and treatment effect heterogeneity assumption is violated. $\Gamma$ simply captures the maximum difference between the conditional ATE based on both observed and unobserved moderators ($\mathbf{X}_i$ and $\mathbf{U}_i$) and the conditional ATE based only on observed moderators ($\mathbf{X}_i$). So, if researchers are worried that observed moderators fail to capture a large amount of treatment effect heterogeneity, they have to choose larger values of $\Gamma$. Even when some moderators are unobserved, if it is plasusible to assume that observed moderators capture the majority of treatment effect heterogeneity, they can choose smaller values of $\Gamma$.

For example, suppose it is plausible to assume that the difference between the conditional ATE based on both observed and unobserved moderators and the conditional ATE based only on observed moderators is at most 3 percentage points and likely to be about 1 percentage point. In this case, researchers can choose, for example, $\Gamma = (0.005, 0.01, 0.03)$, which captures values ranging from 0.5 to 3 percentage points, to investigate plausible scenarios where the ignorability of sampling and treatment effect heterogeneity assumption is violated. Once $\Gamma$ is chosen, solving the minimization problem (equation (12)) is essentially the same as solving the original minimization problem. The only difference is that the threshold is now $\Gamma$ rather than 0.

### A.2.2 Overlap

**Overview**

A hypothetical population chosen by analysts or skeptics might not be comparable to the experimental sample because the overlap assumption (Assumption 2) might be violated, i.e., the hypothetical population of interest contains a subset of units that are not represented in the experimental data. This overlap assumption is essential for any generalization method, and without this assumption, inference about the T-PATE or external validity depends on extrapolation.

In such scenarios, we cannot consistently estimate the true external robustness, but we can still estimate its upper bound. To formally consider this problem, suppose that $\rho \in [0, 1]$ captures the proportion of a hypothetical population where Assumption 2 holds. When $\rho = 1$, Assumption 2 holds for the entire distribution, and thus, results in Section 3.3.1 can be used. We define $\mathsf{T\text{-}PATE}_{\mathtt{out}}$ to be the average causal effect in a subset of a hypothetical population where Assumption 2 is violated, which is an unknown quantity. We then define a sensitivity parameter $\delta \geq 0$ to quantify its range such that $-\delta \leq \mathsf{T\text{-}PATE}_{\mathtt{out}} \leq \delta$.

Using this formalization, we can write down the KL minimization problem in cases where Assumption 2 is violated.

$$\mathrm{KL}_0^{\dagger\dagger}(\rho, \delta) \;\coloneqq \min_{-\delta \leq \mathsf{T\text{-}PATE}_{\mathtt{out}} \leq \delta} \; \min_{\widetilde{\mathcal{P}} \in \mathcal{CP}} \; \mathrm{KL}(\widetilde{\mathcal{P}} || \mathcal{P}_{\mathtt{exp}}) \quad \text{s.t.} \quad \underbrace{\rho \times \mathbb{E}\{\tau_0(\mathbf{X}_i); \widetilde{\mathcal{P}}\} + (1 - \rho) \times \mathsf{T\text{-}PATE}_{\mathtt{out}} \leq 0}_{\text{T-PATE constraint}}$$

where $\rho \times \mathbb{E}\{\tau_0(\mathbf{X}_i); \widetilde{\mathcal{P}}\} + (1 - \rho) \times \mathsf{T\text{-}PATE}_{\mathtt{out}}$ is the T-PATE when Assumption 2 is violated. Therefore, under this setting, the true external robustness is defined as $\xi_0^{\dagger\dagger}(\rho, \delta) \coloneqq 1 - \exp(-\mathrm{KL}_0^{\dagger\dagger}(\rho, \delta))$.

This setup generalizes the KL minimization problem in equation (8) in two ways. First, the T-PATE constraint now considers the T-PATE in cases where Assumption 2 is violated with two sensitivity parameters $\rho$ and $\delta$. Second, because $\mathsf{T\text{-}PATE}_{\mathtt{out}}$ is inherently unobserved, we consider the worst-case, i.e., minimizing the KL-divergence with respect to $\mathsf{T\text{-}PATE}_{\mathtt{out}}$ within the bound specified by a sensitivity parameter $\delta$.

We show that

$$\mathrm{KL}_0^{\dagger\dagger}(\rho, \delta) \;=\; \min_{\widetilde{\mathcal{P}} \in \mathcal{CP}} \; \mathrm{KL}(\widetilde{\mathcal{P}} || \mathcal{P}_{\mathtt{exp}}) \qquad \text{s.t.} \quad \mathbb{E}\{\tau_0(\mathbf{X}_i); \widetilde{\mathcal{P}}\} \leq \frac{1 - \rho}{\rho} \times \delta.$$

Because $\rho, \delta \geq 0$, $\mathrm{KL}_0^{\dagger\dagger}(\rho, \delta) \leq \mathrm{KL}_0$ and thus, $\xi_0^{\dagger\dagger}(\rho, \delta) \leq \xi_0$. The equality holds when $\rho = 1$ or $\delta = 0$. Therefore, the proposed estimator is an estimator for the upper bound of the true external robustness. We provide the proof in the next subsection.

Several points are worth noting. First, when researchers are willing to specify $(\rho, \delta)$, they can explicitly conduct a sensitivity analysis to see how estimated external robustness changes according to $(\rho, \delta)$. Second, as we might expect intuitively, this formalization suggests that when

Assumption 2 is violated for a larger fraction of populations, the true external robustness might be much lower than the upper bound we estimate. Finally, this also shows that when the T-PATE for an unrepresented subset of units is much smaller than 0, the true external robustness is also much lower than the upper bound we estimate. By doing an explicit sensitivity analysis, researchers can examine these concerns in practice.

**Optimization Problem**

When the overlap assumption is violated, we consider the following general KL minimization problem.

$$\begin{aligned} & \mathrm{KL}_0^{\dagger\dagger}(\rho, \delta) \\ := & \min_{-\delta \leq \mathsf{T\text{-}PATE}_{\mathtt{out}} \leq \delta} \min_{\widetilde{\mathcal{P}} \in \mathcal{CP}} \mathrm{KL}(\widetilde{\mathcal{P}} || \mathcal{P}_{\mathtt{exp}}) \quad \text{s.t.} \quad \rho \times \mathbb{E}\{\tau_0(\mathbf{X}_i); \widetilde{\mathcal{P}}\} + (1-\rho) \times \mathsf{T\text{-}PATE}_{\mathtt{out}} \leq 0 \end{aligned} \tag{13}$$

where $\rho \in [0, 1]$ captures the proportion of a hypothetical population where Assumption 2 holds. When $\rho = 1$, Assumption 2 holds for the entire distribution, and thus, results in Section 3.3.1 can be used. We define $\mathsf{T\text{-}PATE}_{\mathtt{out}}$ to be the average causal effects in a subset of a hypothetical population where Assumption 2 is violated, which is an unknown quantity. We then define a sensitivity parameter $\delta \geq 0$ to quantify its range such that $-\delta \leq \mathsf{T\text{-}PATE}_{\mathtt{out}} \leq \delta$. Then, $\rho \times \mathbb{E}\{\tau_0(\mathbf{X}_i); \widetilde{\mathcal{P}}\} + (1-\rho) \times \mathsf{T\text{-}PATE}_{\mathtt{out}}$ is the T-PATE in the hypothetical population of interest.

By rearranging terms, we get

$$\begin{aligned} & \mathrm{KL}_0^{\dagger\dagger}(\rho, \delta) \\ = & \min_{-\delta \leq \mathsf{T\text{-}PATE}_{\mathtt{out}} \leq \delta} \min_{\widetilde{\mathcal{P}} \in \mathcal{CP}} \mathrm{KL}(\widetilde{\mathcal{P}} || \mathcal{P}_{\mathtt{exp}}) \quad \text{s.t.} \quad \mathbb{E}\{\tau_0(\mathbf{X}_i); \widetilde{\mathcal{P}}\} \leq -\frac{1-\rho}{\rho} \times \mathsf{T\text{-}PATE}_{\mathtt{out}} \end{aligned} \tag{14}$$

Given that the objective function is a decreasing function of $-\frac{1-\rho}{\rho} \times \mathsf{T\text{-}PATE}_{\mathtt{out}}$, $\frac{1-\rho}{\rho}$ is positive, and $-\mathsf{T\text{-}PATE}_{\mathtt{out}} \leq \delta$, we have

$$\mathrm{KL}_0^{\dagger\dagger}(\rho, \delta) = \min_{\widetilde{\mathcal{P}} \in \mathcal{CP}} \mathrm{KL}(\widetilde{\mathcal{P}} || \mathcal{P}_{\mathtt{exp}}) \qquad \text{s.t.} \quad \mathbb{E}\{\tau_0(\mathbf{X}_i); \widetilde{\mathcal{P}}\} \leq \frac{1-\rho}{\rho} \times \delta. \tag{15}$$

Because $\rho, \delta \geq 0$, $\mathrm{KL}_0^{\dagger\dagger}(\rho, \delta) \leq \mathrm{KL}_0$ and thus, $\xi_0^{\dagger\dagger}(\rho, \delta) \leq \xi_0$. The equality holds when $\rho = 1$ or $\delta = 0$. Therefore, the proposed estimator is an estimator for the upper bound of the true external robustness.

**Sensitivity Analysis**

As we described for the ignorability of sampling and treatment effect heterogeneity assumption, researchers can conduct an explicit sensitivity analysis to understand how external robustness changes depending on the degree of violation of the assumption.

In particular, researchers can choose different values of $(\rho, \delta)$ to consider a wide range of scenarios where the overlap assumption is violated. First, we consider $\rho \in [0, 1]$, which simply

captures the proportion of a hypothetical population where the overlap assumption holds. So, if researchers are worried that the experimental data fail to capture a large portion of a hypothetical population of interest, they have to choose larger values of $\rho$. Even when the overlap assumption is violated, if it is plausible to assume that the experimental data capture the majority of the hypothetical population of interest, they can choose smaller values of $\rho$. For example, suppose it is plausible to assume that the experimental data captures roughly 80 percent of the hypothetical population of interest. In this case, researchers can choose, for example, $\rho = (0.70, 0.80, 0.90)$, which captures values ranging from 70 to 90 percents, to investigate different plausible scenarios where the overlap assumption is violated.

Researchers also have to choose values of $\delta$. $\delta \geq 0$ simply captures a plausible range of $\mathsf{T\text{-}PATE}_{\mathtt{out}}$, which is the average causal effects in a subset of a hypothetical population where the overlap assumption is violated. Intuitively, this captures uncertainty about the magnitude of $\mathsf{T\text{-}PATE}_{\mathtt{out}}$.

So, if researchers are uncertain about the magnitude of $\mathsf{T\text{-}PATE}_{\mathtt{out}}$, they have to choose larger values of $\delta$. If it is plausible to assume that the magnitude of $\mathsf{T\text{-}PATE}_{\mathtt{out}}$ is small, they can choose smaller values of $\delta$. For example, suppose it is plausible to assume that the magnitude of $\mathsf{T\text{-}PATE}_{\mathtt{out}}$ is about 5 percentage points. In this case, researchers can choose, for example, $\delta = (0.03, 0.05, 0.07)$, which captures values ranging from 3 to 7 percentage points, to investigate different plausible scenarios where the overlap assumption is violated.

Once $\rho$ and $\delta$ are chosen, solving the minimization problem (equation (15)) is essentially the same as solving the original minimization problem. The only difference is that the threshold is now $(1-\rho)/\rho \times \delta$ rather than 0.

## A.3 Incorporating Uncertainties

In this paper, to incorporate uncertainty, we rely on nonparametric bootstrap. Suppose the SATE estimate is positive. Then, we estimate external robustness such that the lower confidence interval of the T-PATE, approximated by nonparametric bootstrap, is equal to or smaller than zero. In particular, we use the following procedure.

First, we solve equation (6) and estimate weights such that the T-PATE is less than or equal to zero. We denote them as $\widehat{w}_p$. Then, we use nonparametric bootstrap to estimate standard errors of the T-PATE estimator. In particular, we resample $n$ observations from the experimental data $(Y_i, T_i, \mathbf{X}_i, \widehat{w}_{pi})_{i=1}^n$ with replacement. With the bootstrap data, we re-estimate the CATE function and then estimate $\widehat{\theta}_b = \frac{1}{n}\sum_{i \in S_b} \widehat{w}_{pi}\widehat{\tau}_b(\mathbf{X}_i)$ where $S_b$ represents a set of unit indices that belong to the $b$th bootstrap data and $\widehat{\tau}_b(\cdot)$ is the CATE function estimated with the $b$th bootstrap data. Then, the standard error is estimated to be the empirical standard deviation of the bootstrap estimates, i.e., $\widehat{\text{se}} = \sqrt{\frac{1}{B}\sum_{b=1}^B (\widehat{\theta}_b - \overline{\widehat{\theta}}_B)^2}$ where $B$ is the total number of bootstap and $\overline{\widehat{\theta}}_B = \frac{1}{B}\sum_{b=1}^B \widehat{\theta}_b$. Finally, when we consider $(1-\alpha)$ confidence interval, we solve

the following modified optimization problem.

$$\begin{aligned}\widehat{\mathrm{KL}}_{1-\alpha} \quad = \quad & \min_{\mathbf{w}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} w_i \log(w_i)}_{\text{KL-divergence}} \\ & \text{s.t.} \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} w_i \widehat{\tau}(\mathbf{X}_i) - \Phi_{1-\alpha/2} \times \widehat{\mathrm{se}} \leq 0,}_{\text{T-PATE constraint}} \quad \underbrace{\sum_{i=1}^{n} w_i = n, \quad w_i \geq 0,}_{\text{Standard weights constraints}} \end{aligned} \tag{16}$$

where $\Phi_\gamma$ is the $\gamma$ quantile of the standard normal distribution. For example, when we consider the approximated 95% confidence interval of the T-PATE, we choose $\Phi_{1-\alpha/2} = 1.96$. The external robustness is defined as $\widehat{\xi}_{1-\alpha} = 1 - \exp(-\widehat{\mathrm{KL}}_{1-\alpha})$. Therefore, in this problem, we estimate external robustness such that the lower confidence interval of the T-PATE, approximated by nonparametric bootstrap, is equal to or smaller than zero. When the SATE estimate is negative, we estimate external robustness such that the upper confidence interval of the T-PATE is equal to or larger than zero.

# B Extensions

## B.1 Incorporating Partial Knowledge about Population Data

We have so far assumed that researchers do not have any information about potential populations. This is a core advantage of our approach as researchers often do not have data on populations in many applications.

However, in some scenarios, we have partial knowledge about population data, and we might want to estimate robustness only against "plausible" populations. For example, suppose education is an important moderator, and we know the "plausible" proportion of college graduates is between 30 and 50 % in relevant populations. In this case, we can estimate external robustness only against populations that have a proportion of college graduates between 30 and 50 %.

Formally, we want to add a new constraint about covariates $\mathbf{Z}_i$ to the problem. Note that researchers can choose covariates $\mathbf{Z}_i$ that are different from $\mathbf{X}_i$ used to estimate CATEs, while in most applications, $\mathbf{Z}_i$ will be a subset of $\mathbf{X}_i$ because researchers tend to have much less information about covariates in populations than in the experimental sample. Below, we consider several constraints we can add to the original minimization problem.

**Example 1** (the marginal means of $\mathbf{Z}$ in populations are known)

$$\frac{1}{n}\sum_{i=1}^{n} w_i \mathbf{Z}_i = \mathbf{z}_{\text{target}}$$

where $\mathbf{z}_{\text{target}}$ is a vector of the marginal means in the target population. For example, we might know the proportion of college graduates is 40% in relevant populations.

**Example 2** (the marginal means of $\mathbf{Z}$ in populations are within some ranges)

$$\mathbf{z}_{\text{target}}^{L} \leq \frac{1}{n}\sum_{i=1}^{n} w_i \mathbf{Z}_i \leq \mathbf{z}_{\text{target}}^{U}$$

where $\mathbf{z}_{\text{target}}^{L}$ and $\mathbf{z}_{\text{target}}^{U}$ are vectors of lower and upper bounds of the marginal means in the target population. For example, we might know the proportion of college graduates is between 30% and 50% in relevant populations.

**Example 3** (the marginal means of $\mathbf{Z}$ in populations are smaller than those in the experiment)

$$\frac{1}{n}\sum_{i=1}^{n} w_i \mathbf{Z}_i \leq \frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}_i$$

where $\frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}_i$ is a vector of the marginal means in the experimental data. For example, we might know that college graduates are overrepresented in the experiment.

### B.1.1 Details

We can incorporate partial knowledge about population data as constraints, and we provide some examples above. To estimate external robustness, we have to estimate how much additional amount of reweighting is required to make the T-PATE less than or equal to zero.

To capture this formally, we use a two-step approach. First, we estimate weights that satisfy constraints implied by partial knowledge about population data without including the T-PATE constraint. Then, we estimate how much additional re-weighting is required when we add the T-PATE constraint.

In particular, in the first step, we solve the following minimization problem to estimate weights.

$$\begin{aligned}
\widehat{\mathbf{w}}_1 \quad = \quad & \underset{\mathbf{w}_1}{\operatorname{argmin}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} w_{i1}\log(w_{i1})}_{\text{KL-divergence}} \\
& \text{s.t.} \quad \underbrace{\mathbf{z}_{\text{target}}^{L} \leq \frac{1}{n}\sum_{i=1}^{n} w_{i1}\mathbf{Z}_i \leq \mathbf{z}_{\text{target}}^{U},}_{\text{Constraints about Populations}} \quad \underbrace{\sum_{i=1}^{n} w_{i1} = n, \quad w_{i1} \geq 0.}_{\text{Standard weights constraints}}
\end{aligned}$$

where we use Example 2 as constraints for population data, but researchers can use any other relevant constraints.

Then, in the second step, we solve the following minimization problem that additionally includes the T-PATE constraint.

$$\widehat{\mathbf{w}}_2 \quad = \quad \underset{\mathbf{w}_2}{\operatorname{argmin}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \widehat{w}_{i1} w_{i2}\log(\widehat{w}_{i1} w_{i2})}_{\text{KL-divergence}}$$

$$\text{s.t.} \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} \widehat{w}_{i1} w_{i2} \widehat{\tau}(\mathbf{X}_i) \leq 0,}_{\text{T-PATE constraint}}$$

$$\underbrace{\mathbf{z}^L_{\text{target}} \leq \frac{1}{n}\sum_{i=1}^{n} \widehat{w}_{i1} w_{i2} \mathbf{Z}_i \leq \mathbf{z}^U_{\text{target}},}_{\text{Constraints about Populations}} \qquad \underbrace{\sum_{i=1}^{n} w_{i2} = n, \quad w_{i2} \geq 0.}_{\text{Standard weights constraints}}$$

Therefore, if the additional T-PATE constraint does not require any additional reweighting, we should get uniform weights as the solution for $\mathbf{w}_2$. Finally, we define a measure of external robustness as the additional amount of reweighting captured by the KL-divergence and its transformation of estimated $\widehat{\mathbf{w}}_2$.

$$\widehat{\xi} \coloneqq 1 - \exp(-\frac{1}{n}\sum_{i=1}^{n} \widehat{w}_{i2} \log(\widehat{w}_{i2})). \tag{17}$$

## B.2 Subgroup Analysis

In many experiments, researchers are often interested in subgroup analyses to investigate causal mechanisms and explain how treatment effects vary across theoretically relevant subgroups, e.g., democrats and republicans. In such scenarios, researchers can evaluate external robustness separately for subgroups. For example, they can evaluate whether causal conclusions for democrats (republicans) in the experiment generalize to other populations of democrats (republicans). Even if causal effects are heterogeneous across democrats and republicans, as long as treatment effects are more homogeneous within each group, causal conclusions for democrats and republicans might still be externally robust.

However, it is important to emphasize the risk of overfitting. If researchers estimate external robustness for many subgroups, they are likely to find large external robustness simply by chance, even if the true external robustness is low. This is similar to the problem of $p$-hacking where researchers explore many subgroups in experiments and selectively report large effects.

Therefore, we follow the standard recommendations for subgroup analyses in randomized experiments. First, researchers should always report external robustness using all observations in an experiment (as we report the overall SATE estimate before showing subgroup analyses in experimental studies). Second, researchers should ideally pre-register subgroup analyses before implementing a randomized experiment, and should justify subgroups theoretically if they choose subgroups at the analysis stage of the experiment. What is important here is that subgroups relevant in the usual subgroup analysis for CATEs are exactly the same as subgroups relevant in the external robustness estimation. Therefore, researchers do not need to additionally pre-register or justify the choice of subgroups for estimating external robustness. Estimation of subgroup external robustness can directly use subgroups that researchers select for estimation of CATEs.

## B.3 External Validity with respect to Contexts

Social scientists are often interested in external validity with respect to not only populations but also contexts, which are called $X$- and $C$-validity, respectively (Egami and Hartman, 2022). We often discuss geography and time as important contexts. For example, researchers might be interested in understanding whether and how we can generalize experimental results found in California to other states. This $C$-validity question is challenging because all observations are from California and the variable "State" has no variation within the experimental data.

One approach to address $C$-validity is to consider contextual moderators $\mathbf{M}_i$ — variables related to mechanisms through which contexts affect outcomes and moderate treatment effects (Egami and Hartman, 2022). If contexts affect outcomes only through measured context-moderators, we can formalize the $C$-validity problem again as a question of reweighting. Thus, we can estimate external robustness with respect to both populations and contexts by including covariates $\mathbf{X}_i$ related to $X$-validity and context-moderators $\mathbf{M}_i$ related to $C$-validity.

### B.3.1 Details

Dedine $C$ to be an indicator variable for contexts where $C = 0$ indicates a context under which the experiment is done. $C = c$ ($c \leq 0$) represents other contexts where no experimental data exist. Formally, the question of $C$-validity can be formally written as follows.

For any populations $\widetilde{\mathcal{P}}$,

$$\mathbb{E}(Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}, C_i = 0; \widetilde{\mathcal{P}}) \;\neq\; \mathbb{E}(Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}, C_i = c; \widetilde{\mathcal{P}}), \tag{18}$$

and, for $\forall c \neq 0$,

$$\Pr(C_i = 0; \mathcal{P}_{\texttt{exp}}) = 1 \quad \text{and} \quad \Pr(C_i = c; \mathcal{P}_{\texttt{exp}}) = 0. \tag{19}$$

One approach to address $C$-validity is to consider contextual moderators $\mathbf{M}_i$ — variables related to mechanisms through which contexts affect outcomes and moderate treatment effects. This idea is formalized as the contextual exclusion restriction in Egami and Hartman (2022). This assumption is equivalent to assuming that there exists contextual moderators $\mathbf{M}_i$ that satisfy the following two conditions.

For any populations $\widetilde{\mathcal{P}}$,

$$\mathbb{E}(Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}, \mathbf{M}_i = \mathbf{m}, C_i = 0; \widetilde{\mathcal{P}}) \;=\; \mathbb{E}(Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}, \mathbf{M}_i = \mathbf{m}, C_i = c; \widetilde{\mathcal{P}}), \tag{20}$$

and, for $\forall \mathbf{m}$,

$$0 < \Pr(\mathbf{M}_i = \mathbf{m}; \mathcal{P}_{\texttt{exp}}). \tag{21}$$

Therefore, under the contextual exclusion restriction, we can estimate external robustness to both $X$- and $C$-validity jointly by considering the conditional ATE $\mathbb{E}(Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}, \mathbf{M}_i = \mathbf{m}, C_i = 0; \mathcal{P}_{\texttt{exp}})$. In practice, researchers just need to include covariates $(\mathbf{X}_i, \mathbf{M}_i)$ rather than $\mathbf{X}_i$ alone when estimating external robustness.

Similar to the case of the ignorability of sampling and treatment effect heterogeneity, we can derive similar results — the proposed estimator is the upper bound of the true external robustness when contextual exclusion restriction is violated.

We note that there potentially exist other approaches to estimating external robustness that take into account $C$-validity, and future work can explore such alternative approaches.

### B.4 Observational Studies

External validity is equally important for experimental and observational studies. However, for observational studies, we also have to address internal validity concerns. In this subsection, we consider the most standard approach using the conditional ignorability assumption. In the experimental data distribution $\mathcal{P}_{\texttt{exp}}$,

$$Y_i(1), Y_i(0) \perp\!\!\!\perp T_i \mid \mathbf{X}_i. \tag{22}$$

Under this assumption, we can identify the conditional ATE function as in the experimental data. Thus, researchers can use the same proposed method to estimate external robustness.

Future work can also explore how to extend the proposed method to other quasi-experimental methods (e.g., instrumental variable methods and difference-in-differences). Because these other quasi-experimental methods cannot identify the sample average treatment effect without additional assumptions (e.g., instrumental variable methods focus on the complier average treatment effect and difference-in-differences on the average treatment effect on the treated) even when they only focus on internal validity, substantial development is required to properly define and estimate external robustness, which is beyond the scope of this paper.

## C Benchmarks

In this section, we discuss in detail the process of creating benchmarks for external robustness.

Section 3 formally introduces how external robustness quantifies the minimal amount of reweighting required to explain away the T-PATE. This metric is expressed as a number between 0 and 1. To use this number, researchers must have an understanding of "how much reweighting is large." One way to answer this question is to compare any given value of external robustness to benchmarks that can be substantively interpreted based on pre-existing knowledge.

One of our main contributions is to construct such benchmarks: to do so, we rely on two widely used types of surveys: national surveys and samples collected using Amazon Mechanical Turk (MTurk). We use variation in the reweighting required for each of these survey types to approximate representativeness to identify two benchmark values: one corresponding to a "small" amount of reweighting and one corresponding to a "moderate" amount of reweighting.

### C.1 Process

The representativeness of a survey is the degree to which the survey sample resembles the population of interest. In a more representative survey, the distribution of covariate values is

closer to the distribution of covariate values in the population of interest. When the sample is randomly drawn from the entire population of interest, these distributions are identical in expectation. In practice, some selection bias often implies that the sampling frame is distinct from the population of interest. In such cases, representativeness of the sample can be approximated by reweighting sample observations such that the moments of the reweighted sample are equal to those of the population of interest. When selection bias is smaller, the amount of reweighting required to approximate the population of interest is smaller.

The amount of selection bias is limited in national surveys for which researchers invest a significant amount of effort at the design stage to ensure that the sample resembles the national population. We therefore understand the amount of reweighting required for these sample to be "small".

By contrast, MTurk online samples result from the survey of a set of volunteer "crowdworkers" that are likely to differ from the general population by virtue of having voluntarily selected into paid MTurk tasks. The reweighting required to approximate the population of interest is therefore larger. Common findings in the political science literature are that MTurk samples are often more representative of the U.S. population than in-person convenience samples, but less representative than national probability surveys (Berinsky, Huber and Lenz, 2012). We therefore understand the associated amount of reweighting to be "moderate".

To express different amounts of reweighting on the same scale as our measure of external robustness, we quantify them using the procedure introduced in Section 3. That is, given a vector of weights $\widehat{\mathbf{w}}_s$, we compute $R_s = 1 - \exp(-\mathrm{KL}(\widehat{\mathbf{w}}_s))$, where $\mathrm{KL}(\cdot)$ is the Kullback–Leibler divergence.

In the next two sections, we describe how the surveys are selected and how the weights are obtained for each survey type.

## C.2 National Surveys

In this section, we describe how we assembled the data used to construct the benchmark based on national surveys. In particular, we explain how we chose what surveys to use, what waves to include, and what weights to use.

### C.2.1 Surveys

We selected eight national surveys: two surveys from the United States (ANES and CCES/CES), two surveys from Europe (ESS and Eurobarometer), and four surveys from the Global Barometer Surveys (GBS) Project (Afrobarometer, Latinobarometro, Arab Barometer, and Asian Barometer Survey). These surveys are all conducted by academic or public institutions, and widely used in the political science literature to study their regions of application. We did not include the other existing survey from the GBS Project (Eurasia Barometer) because the data was not publicly available at the time of writing.

All of these surveys are meant to reflect the corresponding regional populations. To this

end, researchers stratify the population (at least by region, sometimes by other demographic characteristics), and then assign an equal probability of being drawn to each adult (or eligible voter in the case of the ANES) in each stratum. Weights required to collapse the remaining differences between the sample and the populations of interest are then directly provided by the survey administrators.

### C.2.2 Waves and Weights

We selected five waves from each survey, as described below. When more than one weight was provided by the researchers, we selected one of them as recommended by the accompanying documentation and in a way that was as consistent as possible across waves of a given survey.

- **American National Election Survey (ANES)**: we used the last five available waves of the ANES: 2004, 2008, 2012, 2016, and 2020. The ANES is currently implemented as a collaboration between Stanford University and the University of Michigan, and is funded by the National Science Foundation. As weights provided by the researchers vary from one wave to the next, we detail the data and weights used in each iteration below:
  - **ANES 2004**: we used the pre-election survey, and accordingly used V040101 as the weight.
  - **ANES 2008**: we used the pre-election survey, and accordingly used V080101 as the weight.
  - **ANES 2012**: we used the full sample, including both face-to-face and internet respondents, and accordingly used WEIGHT_FULL variable as the weight. No survey sub-sample and weight equivalent to the other four waves was available.
  - **ANES 2016**: we used the pre-election survey, including both face-to-face and internet respondents, and accordingly used V160101 as the weight.
  - **ANES 2020**: we used the pre-election survey and accordingly used V200010A as the weight.
- **Cooperative Election Study (CES), formerly the Cooperative Congressional Election Study (CCES)**: we used five two year-distanced waves of the CCES/CES: 2012, 2014, 2016, 2018, and 2020. More recent waves were available in 2019 and 2017 but we decided to include only election years, as these typically include more observations. The CES is currently implemented by Harvard University and YouGov and funded by the National Science Foundation.
  - **CCES 2012**: we used the full survey with WEIGHT_VV (voter-validated team weights) as the weight.

- **CCES 2014**: we used the full survey with WEIGHT (common weights) as the weight. No voter-validated weight was provided for this wave.
- **CCES 2016**: we used the full survey with COMMONWEIGHT_VV (voter-validated common-content weights) as the weight.
- **CCES 2018**: we used the full survey with VVWEIGHT (voter-validated weights) as the weight.
- **CCES 2020**: we used the full survey with VVWEIGHT (voter-validated weights) as the weight.

- **European Social Survey**: we used the last five available rounds of the ESS: 2010, 2012, 2014, 2006, and 2018. For all waves, we used the full sample, with PSPWGHT (Post-stratification weight including design weight) as the weight. The ESS is implemented by the European Research Infrastructure Consortium, and funded by the 25 participating country in proportion to their respective GDP.

- **Eurobarometer**: we used the first round in the last five Standard Eurobarometer surveys distanced by at least a year: 88 (2017), 90 (2018), 92 (2019), 94 (2020), and 95 (2021). We therefore did not include the waves 89, 91, and 93, respectively conducted in the same years as waves 90, 92, and 94. For all waves, we used the full sample with the post-stratification weights W1 as the weight. The Eurobarometer is the polling instrument of the European Commission.

- **Afrobarometer**: we used the last five waves of the Afrobarometer: 3 (2015), 4 (2008), 5 (2011-2013), 6 (2016), and 7 (2019). For each wave, we used COMBINWT (combined weights within and across countries) as the weight. The Afrobarometer is a pan-African nonprofit, a member of the GBS Project, and is supported by Michigan State University (MSU) and the University of Cape Town (UCT).

- **Latinobarometro**: we used the last five waves of the Latinobarometro: 2015, 2016, 2017, 2018, and 2020. For each wave, we used the within-country poststratification weights WT as the weight. Because weights were not available for all countries, we computed the amount of reweighting required ($R_s$) within each country and then averaged that value over all the countries for which weights were provided. The Latinobarometro is a Chilean nonprofit, a member of the GBS Project, and its advisory board includes a wide group of academics.

- **Arab Barometer**: we used the last five waves of the Arab Barometer: II (2006-2009), III (2010-2011), IV (2012-2014), V (2016-2017), and VI (2017-2018). For each wave, we used the within-country poststratification weights WT as the weight. Because weights were not available for all countries, we computed the amount of reweighting required

($R_s$) within each country and then averaged that value over all the countries for which weights were provided. The Arab Barometer is a member of the GBS Project, and its steering committee and principal investigators include academics at the Center for Strategic Studies at the University of Jordan in Amman, the Palestinian Center for Policy and Survey Research in Ramallah, the Social and Economic Survey Research Institute at Qatar University in Doha, Princeton University, and the University of Michigan.

- **Asian Barometer Survey**: we used all five waves of the Asian Barometer: 1 (2001-2003), 2 (2005-2008), 3 (2010-2012), 4 (2014-2016), and 5 (2018-2021). For waves 1 to 4, we used the following cross-country weights: W_ALL for waves 1 and 2, COUWEIGHT for wave 3, and W_CROSS for wave 4. For wave 5, the data was not yet aggregated by the researchers. We used the within-country weights W_ALL, computed the amount of reweighting required ($R_s$) within each country, and then averaged that value over all the available countries. The Asian Barometer Survey is a member of the GBS Project and is co-hosted by the Institute of Political Science, Academia Sinica and the Institute for the Advanced Studies of Humanities and Social Sciences, National Taiwan University.

For each survey $s$, we computed the amount of reweighting $R_s$ required to approximate the population of interest using the samples and weights described above. Table 1 displays this amount, computed on a scale from 0 to 1 as described in section 3, for all national surveys.

## C.3 Mechanical Turk Samples

We used replication data from four published research articles from the political science literature relying on samples collected on MTurk: Berinsky et al (2012), Huff and Tingley (2015), Mullinix et al. (2015), and Coppock et al. (2018). All four studies explicitly study the representativity and/or usability of MTurk online samples, have been widely cited, and provide replication data.

### C.3.1 Estimation of weights

In contrast to the national surveys, no weights systematically accompany the MTurk surveys. Computing $R_s$, the amount of reweighting required to approximate the population of interest, therefore first requires estimating these weights ourselves. For each study, we followed the process outlined below:

1. **Selection of a corresponding representative sample**: we want to reweight the MTurk observations such that the distribution of covariate values after reweighting is identical to this representative survey. All articles included such representative samples, as detailed in the next section, and used them as a basis for comparison with the MTurk data.

Table 1: Amount of Reweighting $R_s$ by National Survey

| Survey | Year | $R_s$ |
|---|---|---|
| ANES | 2020 | 0.31 |
| ANES | 2016 | 0.17 |
| ANES | 2012 | 0.29 |
| ANES | 2008 | 0.23 |
| ANES | 2004 | 0.09 |
| CES | 2020 | 0.27 |
| CCES | 2018 | 0.28 |
| CCES | 2016 | 0.27 |
| CCES | 2014 | 0.34 |
| CCES | 2012 | 0.35 |
| ESS | 2018 | 0.12 |
| ESS | 2016 | 0.12 |
| ESS | 2014 | 0.09 |
| ESS | 2012 | 0.12 |
| ESS | 2010 | 0.13 |
| Eurobarometer | 2021 | 0.09 |
| Eurobarometer | 2020 | 0.12 |
| Eurobarometer | 2019 | 0.09 |
| Eurobarometer | 2018 | 0.10 |
| Eurobarometer | 2017 | 0.08 |
| Afrobarometer | 2019 | 0.11 |
| Afrobarometer | 2016 | 0.14 |
| Afrobarometer | 2011-2013 | 0.13 |
| Afrobarometer | 2008 | 0.15 |
| Afrobarometer | 2005 | 0.06 |
| Latinobarometro | 2020 | 0.10 |
| Latinobarometro | 2018 | 0.08 |
| Latinobarometro | 2017 | 0.09 |
| Latinobarometro | 2016 | 0.08 |
| Latinobarometro | 2015 | 0.13 |
| Arab barometer | 2017-2018 | 0.16 |
| Arab barometer | 2016-2017 | 0.06 |
| Arab barometer | 2012-2014 | 0.15 |
| Arab barometer | 2010-2011 | 0.09 |
| Arab barometer | 2006-2009 | 0.20 |
| Asian barometer survey | 2018-2021 | 0.09 |
| Asian barometer survey | 2014-2016 | 0.14 |
| Asian barometer survey | 2010-2012 | 0.12 |
| Asian barometer survey | 2005-2008 | 0.09 |
| Asian barometer survey | 2001-2003 | 0.10 |

2. **Selection of covariates**: we selected the largest set of basic demographic covariates common to both the MTurk sample and the corresponding representative sample. Within 'basic' demographic covariates, we include: age, gender, race, income, education, region, and partisan identity. No interaction between the selected covariates was included.

3. **Computation of weights**: the vector of weights was calculated using the raking procedure (Deville and Särndal, 1992), which was implemented by using the `calibrate()` function of the `survey` package in `R`. The resulting mean of the covariates selected above should be the same for the *reweighted* MTurk sample and the *unweighted* corresponding representative sample.

#### C.3.2 Representative sample and covariates by study

- **Berinsky et al. (2012)**: this article provided two accompanying representative samples: one from the 2008 ANES, and one from the 2008 Current Population Survey (CPS). The MTurk data was collected in 2010. We estimated the weights twice, using each representative sample, and include both results in our final construction of a benchmark for MTurk samples. The covariates included are age, gender, race, and education for both surveys, as well as income for CPS.

- **Huff and Tingley (2015)**: the representative sample provided is from the 2012 CCES. The MTurk data was collected at the same as the CCES in the fall of 2012. The covariates included are age, gender, race, and education.

- **Mullinix et al. (2015)**: we used study 2 from Mullinix et al. (2015), for which the representative sample provided were several iterations of the National Science Foundation funded Time-sharing Experiments for the Social Sciences (TESS). The MTurk data was collected between 2011 and 2013 and its original aim was to replicate the TESS experiments. The covariates included are age, gender, race, and education.

- **Coppock et al. (2018)**: this study provided 27 distinct MTurk samples, and 27 corresponding representative samples. The representative samples were drawn from TESS—some of them were the same experiments used in Mullinix et al (2015)—and the rest was drawn from GfK's KnowledgePanel. The aim of the article was also to study the replicability of TESS and KnowledgePanel results with MTurk online samples. Note that whereas Mullinix et al. (2015) collected one MTurk sample of 14,221 respondents and used it to replicate all experiments, Coppock et al. (2018) collected 27 distinct samples, for a total of 101,745 individual survey responses. Available covariates varied depending on the samples, but included between 3 and 6 of the following 6 variables: age, gender, race, education, and two variables capturing partisan identity/ideology. All samples included age and race, and all but one included gender.

For each MTurk sample $s$, after estimating the weights following the procedure outlined above, we computed the amount of reweighting $R_s$ required to approximate the corresponding representative sample. Table 2 displays this amount, computed on a scale from 0 to 1 as described in Section 3.

Table 2: Amount of Reweighting $R_s$ by MTurk sample

| Source | Representative sample | $R_s$ |
|---|---|---|
| Berinsky et al. (2012) | ANES | 0.70 |
| Berinsky et al. (2012) | CPS | 0.67 |
| Huff and Tingley (2015) | CCES | 0.76 |
| Mullinix et al. (2015) | TESS | 0.65 |
| Bergan (2012) and Coppock et al. (2018) | GFK | 0.65 |
| Brandt (2013) and Coppock et al. (2018) | GFK | 0.62 |
| Caprariello and Reis (2013) and Coppock et al. (2018) | GFK | 0.61 |
| Epley et al. (2009) and Coppock et al. (2018) | GFK | 0.49 |
| Peffley and Hurwitz (2007) and Coppock et al. (2018) | TESS | 0.48 |
| Denny (2012) and Coppock et al. (2018) | GFK | 0.51 |
| Nicholson (2012) and Coppock et al. (2018) | TESS | 0.55 |
| Johnston and Ballard (2016) and Coppock et al. (2018) | TESS | 0.56 |
| Flavin (2011) and Coppock et al. (2018) | GFK | 0.65 |
| Hopkins and Mummolo (2017) and Coppock et al. (2018) | TESS | 0.44 |
| Hiscox (2006) and Coppock et al. (2018) | TESS | 0.38 |
| Gash and Murakami (2009) and Coppock et al. (2018) | GFK | 0.60 |
| Brader (2005) and Coppock et al. (2018) | TESS | 0.58 |
| Jacobsen, Snyder and Saultz (2014) and Coppock et al. (2018) | GFK | 0.58 |
| Murtagh et al. (2012) and Coppock et al. (2018) | GFK | 0.50 |
| McGinty, Webster and Barry (2013) and Coppock et al. (2018) | TESS | 0.61 |
| Parmer (2011) and Coppock et al. (2018) | GFK | 0.57 |
| Chong and Druckman (2010) and Coppock et al. (2018) | TESS | 0.60 |
| Pedulla (2014) and Coppock et al. (2018) | GFK | 0.66 |
| Piazza (2015) and Coppock et al. (2018) | GFK | 0.58 |
| Levendusky and Malhotra (2015) and Coppock et al. (2018) | TESS | 0.41 |
| Shafer (2017) and Coppock et al. (2018) | GFK | 0.60 |
| Transue (2007) and Coppock et al. (2018) | TESS | 0.14 |
| Craig and Richeson (2014) and Coppock et al. (2018) | TESS | 0.58 |
| Thompson and Schlehofer (2014) and Coppock et al. (2018) | GFK | 0.57 |
| Turaga (2010) and Coppock et al. (2018) | GFK | 0.58 |
| Wallace (2011) and Coppock et al. (2018) | GFK | 0.65 |

# D Additional Empirical Application based on a Survey Experiment

Johnston and Ballard (2016) conducted a survey experiment to study how information about expert consensus affects public opinion. In this section, we reanalyze this priming survey experiment about the expert consensus originally conducted by Johnston and Ballard (2016) and replicated by Coppock, Leeper and Mullinix (2018). We use this experiment as both illustration and validation of our proposed approach.

This experiment is uniquely suited for validation because it is conducted both in a national survey (Time-Sharing Experiments for the Social Sciences; TESS) and in MTurk. Therefore, we first mimic a typical situation of survey experimental data analysis by estimating external robustness of this experiment conducted in MTurk as if we did not have access to the TESS data. Then, we can validate our proposed approach by comparing external robustness estimated from the MTurk and causal effects estimated from the TESS data.

We emphasize that we use data from both the MTurk sample and the national survey here only because we want to have a known estimate of the T-PATE to validate our approach. As clarified in Section 5, in practice, our proposed approach of external robustness can be applied (and is most useful) when researchers only have the experimental data and they do not have separate population data.

## D.1 Background

How can scientific experts affect public opinion? In particular, how do citizens respond to consensus about policies among economists? To answer this question, Johnston and Ballard (2016) designed a priming survey experiment, which is a type of experiments that are increasing more popular in the social sciences.

To mimic a typical situation of survey experimental data analysis, we only analyze the experimental data collected in MTurk. The use of online convenience samples (e.g., MTurk and Lucid) and other non-probability samples has become more popular as the cost is lower and researchers can collect data more quickly. Common concerns among users and skeptics are whether causal conclusions found in such convenience samples generalize to other populations, such as the U.S. general population. In this reanalysis, we estimate external robustness to investigate how robust the MTurk experimental results are to external validity bias.

In this Expert Consensus experiment, $2,985$ survey respondents were randomly assigned to treatment and control groups. In the control group, respondents received a statement about one of five policies. In this reanalysis, we focus on a policy on the gold standard because results about external robustness are similar in the other four policies. For completeness, we also report analyses of the other four policies in Appendix E.2. The statement of the gold standard policy is "If the US replaced its discretionary monetary policy regime with a gold standard, defining

a 'dollar' as a specific number of ounces of gold, the price-stability and employment outcomes would be better for the average American." Then, each respondent was asked, "To what extent do you agree or disagree with the following statement?" with response options "Strongly Agree," "Agree," "Disagree," "Strongly Disagree," and "Uncertain." In the treatment group, survey respondents received policy statements identical to those in the control conditions, but each statement was prefaced by the following cue: "A sample of professional economists with widely varying political preferences was asked whether they agreed or disagreed with the following statement." Then, respondents were shown the distribution of responses from professional economists, collected from the Initiative on Global Markets' panel of economists. This panel provides an on-going survey of opinion among prominent economists on salient economic issues. The policies in this study were chosen such that more than 90% of economists share agreement or disagreement with a given statement and there exist no conflicting opinions. For example, for the gold standard policy, 66% of economists strongly disagreed, and 34% disagreed with this statement. After being shown the cue about the expert consensus, each respondent was asked to indicate whether they agreed with the statement, as in the control group. Please see Table 1 of the original paper for more details.

The main outcome variable is whether survey respondents agree with the expert consensus. An estimate of the SATE based on the difference-in-means is 25.4 percentage points (95% CI = $[22.0, 28.8]$; p-value = 0.00). Within the experiment, there is clear evidence of a positive causal effect on average.

## D.2 External Robustness

To estimate external robustness, we rely on all covariates measured in the experimental data. In particular, we use age, education, gender, ideology, partisanship, and race, which are variables typically collected in survey experiments.

Using the chosen variables, we estimate external robustness using the MTurk experimental data alone. To estimate the CATEs as the intermediate step, we again rely on causal forest (Wager and Athey, 2018). Figure 1 shows the estimated external robustness and the distribution of estimated CATEs. The estimated external robustness is 1.00 because estimated CATEs are all positive. Even when we consider the lower limit of the 95% confidence interval of the T-PATE, which is approximated by bootstrap, we found the estimated external robustness to be 1.00. Therefore, this implies the T-PATE estimate is robust to different kinds of populations even when they are substantially different from the experimental sample.

This estimation of external robustness using the MTurk sample suggests that causal effects in the U.S. general population are also likely to be positive. In the next section, we validate this statement by using the same Expert Consensus experiment conducted in the national probability survey.

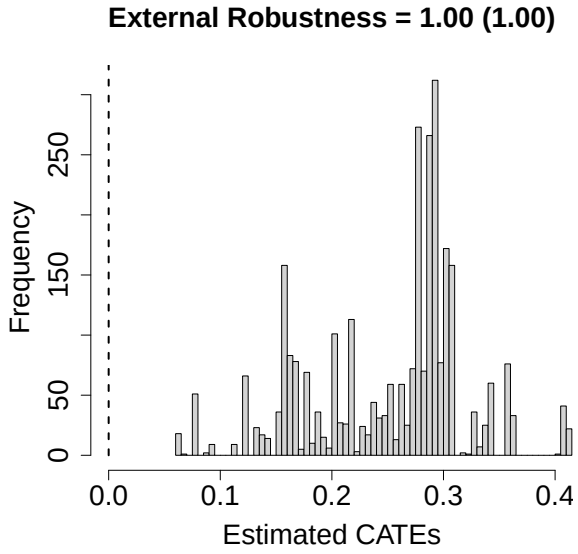



Figure 1: External Robustness and Distribution of Estimated CATEs for the Expert Consensus Experiment conducted in the MTurk. *Note:* The estimated external robustness is 1.00. The estimated external robustness that incorporates uncertainties is also 1.00, reported within parentheses.

### D.3 Empirical Validation

We now empirically validate whether external robustness estimated only using the MTurk experiment can accurately capture the similarity of causal estimates in the MTurk samples and those in the national probability survey. In the national probability survey (the TESS data), the estimated T-PATE is 18.3 percentage points (95% CI = $[10.0, 26.6]$; p-value = 0.00). As estimated external robustness predicted, the SATE and T-PATE estimates have the same sign, while the magnitude of the T-PATE is smaller. This validation shows that researchers can accurately assess whether causal conclusions in the MTurk sample are robust to other populations (including national representative populations) by estimating external robustness only with the experimental data. In Appendix F, we provide a more comprehensive validation study using 13 additional experiments replicated in Coppock, Leeper and Mullinix (2018).

## E Supplementary Information for Empirical Applications

### E.1 Details of Study Design by Domurat, Menashe and Yin (2021)

In this section, we describe some key aspects of study design by Domurat, Menashe and Yin (2021). For further details, readers should refer to sections I and II of the original paper.

The Affordable Care Act (ACA) established regulated insurance marketplaces for individuals without health insurance coverage provided by their employers or by another public program such as Medicaid or Medicare. The marketplace established in California as part of ACA is

called "Covered California" (or CC). Every year, during an "open enrollment" period, individuals can apply to CC to have their eligibility determined, and, if applicable, enroll in a plan. This period lasts from mid-December of the year before to mid-January of the year in question.

### E.1.1 Sampling Population

The sampling population in Domurat, Menashe and Yin (2021) corresponds to the "Funnel" into CC during open enrollment for 2016 coverage (from mid-December 2015 to mid-January 2016). This funnel includes two sets of individuals, for whom an "active determination of eligibility" for CC 2016 coverage was made, but who have not selected a plan yet. These two sets of individuals are the following:

- **Open enrollment applicants**: this set includes individuals who applied to CC during the open enrollment period for 2016 coverage and were determined eligible to enroll.[1]
- **County referrals**: this set includes prior enrollees from the state Medicaid program for whom eligibility to Medicaid changed, and who became eligible to enroll in a plan through CC.

People excluded from the sampling frame (the funnel) therefore include all individuals already covered by CC as well all individuals ineligible to CC—whether or not they applied during open enrollment. In addition, the set of individuals who are potentially eligible but did not receive active determination of eligibility in 2016 are excluded. These include individuals who have been eligible to CC since its establishment, or individuals newly eligible because of a change of eligibility to another plan from a large employer or another public program, with the exception of Medicaid (see "county referrals" above).

### E.1.2 Randomization

The number of individuals in this sampling population (the 2016 Funnel) is not specified in Domurat, Menashe and Yin (2021). 126,182 households were randomly selected from the Funnel. At this sampling stage, randomization was conducted and each household was assigned to one of the five study arms, as described in section IID of the original paper. The researchers then became aware that some of the 126,182 households were not eligible to enroll in CC, or had invalid values for other variables. These households were removed for the final sample of 87,394 households, or 121,828 individuals.

[1]Information on open enrollment applications can be found on the Covered California webpage: `https://www.healthforcalifornia.com/covered-california-enrollment/deadline-and-effective-dates?gclid=EAIaIQobChMIsPCnoKPd-AIVx4xoCR0mbA9UEAAYASAAEgKD5_D_BwE`.

### E.1.3 Table 1 of Domurat, Menashe and Yin (2021) and External Robustness

Table 1 of the original paper contrasts summaries of demographic characteristics across three populations: the sample, CC enrollees, and the general uninsured population of California in 2015, based on data from the American Community Survey (ACS).

Note that the study sample (column 1 of table 1) is a subset of the uninsured population of California (column 3 of table 1). As noted above, column 3 also includes the set of individuals who are potentially eligible but did not receive active determination of eligibility in 2016 are excluded. Column 3 also possibly includes individuals that are not eligible to CC but are still uninsured. For example, an individual eligible to Medicaid who does not take it up remains uninsured. It is not possible from the ACS data alone to determine how many potential CC applicants remain out of the funnel.

Within the set of uninsured individuals potentially eligible to CC—itself an unspecified subset of column 3—it is possible for a systematic relationship between inclusion into the funnel (column 1) and population characteristics to exist, which shows the need for external robustness analysis. If the analysis reveals low external robustness, the results are unlikely to generalize to the set of uninsured individuals who have not received active determination of eligibility into CC because the two populations differ.

## E.2 Additional Analyses for Johnston and Ballard (2016)

In Section D, we focused on the gold standard policy because results for external robustness are similar in the other four policies. For completeness, here we report results on the other four policies.

Table 3 shows results. The first three columns report point estimates, 95% confidence intervals, and p-values for the SATE in the MTurk samples. The next two columns report estimated external robustness and estimated external robustness that incorporates uncertainties. The final three columns point estimates, 95% confidence intervals, and p-values for the T-PATE in the national surveys (the TESS data).

Several points are worth noting. First, as in Section D, we found that estimated external robustness is high for all five policies. Second, again as in Section D, we found that estimated external robustness accurately predicts the similarity of the sign of causal estimates in the MTurk samples and the national surveys, while the T-PATE for the "Immigration" policy is not statistically significant. Please see Appendix F for additional validation.

# F Empirical Validation Study

In Section D, we focused on Johnston and Ballard (2016) to illustrate the use of external robustness. We used one pair of original and replication experiments to validate the performance of external robustness.

In this section, we provide a more comprehensive validation study using all the other ex-

| Outcomes | MTurk Samples | | | External Robustness | | National Probability Surveys | | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | 95% CI | p-value | Estimate | With Uncertainty | Estimate | 95% CI | p-value |
| Gold Standard | 0.25 | (0.22,0.28) | 0.00 | 1.00 | 1.00 | 0.18 | (0.10,0.26) | 0.00 |
| Tax | 0.22 | (0.19,0.25) | 0.00 | 1.00 | 0.97 | 0.18 | (0.10,0.26) | 0.00 |
| Medicare | 0.21 | (0.18,0.24) | 0.00 | 1.00 | 1.00 | 0.11 | (0.02,0.2) | 0.01 |
| China | 0.17 | (0.14,0.20) | 0.00 | 0.98 | 0.89 | 0.10 | (0.02,0.18) | 0.01 |
| Immigration | 0.17 | (0.14,0.20) | 0.00 | 1.00 | 0.99 | 0.06 | (-0.02,0.14) | 0.06 |

Table 3: Results on the other four policies.

periments collected by Coppock, Leeper and Mullinix (2018). In particular, Coppock, Leeper and Mullinix (2018) collected 27 pairs of original and replication experiments, and this set of experiments is uniquely suited for validating our approach because they conducted the same experiments using national surveys (e.g., Time-Sharing Experiments for the Social Sciences; TESS) and the MTurk samples. Therefore, we can validate whether external robustness estimated solely based on the MTurk samples can accurately capture the similarity of causal estimates in the MTurk samples and in national surveys.

The key is that we only use the MTurk samples to estimate external robustness. Therefore, if estimated external robustness captures the similarity of causal estimates in the MTurk samples and in national surveys well, this means that researchers can in practice measure the degree of external robustness reliably just using the experimental data. Indeed, we find that estimated external robustness is high when the SATE estimates in the MTurk samples and the T-PATE estimates in the national surveys have the same sign.

## F.1 Background

To evaluate whether causal findings in the MTurk samples are generalizable to respondents in the national surveys, Coppock, Leeper and Mullinix (2018) collected 27 pairs of experiments; the original experiments were done with national surveys (mostly using TESS; please see the original paper for details of each experiment), and Coppock, Leeper and Mullinix (2018) replicated each experiment with the MTurk samples using the same experimental protocol.

Using this collection of experiments, we empirically validate whether external robustness estimated only using the MTurk experiment can accurately capture the similarity of causal estimates in the MTurk samples and in experiments in the national probability surveys.

We emphasize that this validation is "one-sided" in the sense that we can only validate whether external robustness is correctly high when results in the MTurk samples are generalizable, and we cannot check whether external robustness is correctly low when results in the MTurk samples are not generalizable. This is because the set of experiments collected by Coppock, Leeper and Mullinix (2018) found that most MTurk samples are generalizable, and they didn't find any pair where causal estimates in the MTurk samples and TESS samples have different signs and are both statistically significant. While it is rare to have many pairs of

| Outcomes | MTurk Samples | | | External Robustness | | National Probability Surveys | | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | 95% CI | p-value | Estimate | With Uncertainty | Estimate | 95% CI | p-value |
| Caprariello and Reis (2013) | -0.13 | (-0.24,-0.02) | 0.01 | 0.44 | 0.02 | -0.27 | (-0.48,-0.06) | 0.01 |
| Peffley and Hurwitz (2007) | -0.10 | (-0.19,-0.01) | 0.01 | 0.49 | 0.01 | -0.21 | (-0.33,-0.09) | 0.00 |
| Levendusky and Malhotra (2015) | 0.35 | (0.25,0.45) | 0.00 | 0.86 | 0.61 | 0.10 | (-0.02,0.22) | 0.05 |
| Hiscox (2006) | 0.12 | (0.09,0.15) | 0.00 | 0.99 | 1.00 | 0.22 | (0.13,0.31) | 0.00 |
| Chong and Druckman (2010) | 0.59 | (0.49,0.69) | 0.00 | 1.00 | 1.00 | 0.81 | (0.69,0.93) | 0.00 |
| Hopkins and Mummolo (2017) | -0.21 | (-0.28,-0.14) | 0.00 | 1.00 | 1.00 | -0.13 | (-0.19,-0.07) | 0.00 |
| Brader (2005) | 0.33 | (0.22,0.44) | 0.00 | 1.00 | 0.73 | 0.43 | (0.19,0.67) | 0.00 |
| Brandt (2013) | -0.77 | (-0.86,-0.68) | 0.00 | 1.00 | 1.00 | -0.45 | (-0.61,-0.29) | 0.00 |
| Flavin (2011) | -0.44 | (-0.56,-0.32) | 0.00 | 1.00 | 0.86 | -0.75 | (-0.89,-0.61) | 0.00 |
| Gash and Murakami (2009) | -0.68 | (-0.78,-0.58) | 0.00 | 1.00 | 1.00 | -0.66 | (-0.84,-0.48) | 0.00 |
| Jacobsen, Snyder and Saultz (2014) | 0.47 | (0.37,0.57) | 0.00 | 1.00 | 1.00 | 0.51 | (0.35,0.67) | 0.00 |
| Murtagh et al. (2012) | -0.42 | (-0.61,-0.23) | 0.00 | 1.00 | 0.95 | -0.47 | (-0.72,-0.22) | 0.00 |
| Wallace (2011) | -1.13 | (-1.32,-0.94) | 0.00 | 1.00 | 1.00 | -0.70 | (-0.88,-0.52) | 0.00 |

Table 4: Empirical Validation of External Robustness.

experiments that use the same experimental protocol, if there are pairs of original and replication experiments that have different causal conclusions, future work can use them to further improve empirical validation in this paper.

In particular, we focus on 13 pairs of experiments where the SATE estimates in the MTurk samples and the T-PATE estimates in the national surveys are statistically significant at a conventional level of 0.05. We also removed Johnston and Ballard (2016), which we analyzed in Section D. We check whether external robustness estimated solely based on the MTurk samples are correctly high when causal estimates in the MTurk samples and in the national probability surveys have the same sign. We removed the other 13 pairs because the SATE estimates or the T-PATE estimates are not statistically significant in these studies. In such cases, empirical validation is unreliable because it is difficult to empirically check whether the signs of the SATE and T-PATE estimates are the same.

## F.2 Estimating External Robustness

For each pair of experiments, we use covariates measured in Coppock, Leeper and Mullinix (2018) to estimate conditional ATEs and then estimate external robustness. For most experiments, they measured age, education, gender, ideology, partisanship, and race. To estimate external robustness of the MTurk experiments, we rely on causal forest (Wager and Athey, 2018) to estimate conditional ATEs in each experiment.

Table 4 shows the results. Several points are worth emphasizing. First, 11 out of 13 studies (highlighted in gray) have estimated external robustness larger than 0.57 (the benchmark for "moderate" reweighting). When we look at causal estimates in the national probability surveys for these studies, they in fact have the same sign as the SATE estimates in the MTurk samples. This suggests that, for these studies, by estimating external robustness with the experimental data, researchers can accurately infer the sign of the T-PATE is externally robust. We note that, even though estimated external robustness can infer that the sign of the T-PATE, conducting

experiments in the national probability surveys is still valuable because it helps to learn the exact magnitude of the T-PATE in the target populations.

Finally, for the first two studies on Table 4, estimated external robustness is lower, indicating that they are not externally robust. However, when we look at causal estimates in the national surveys, they have the same sign as those of the SATE estimate, and the magnitude of causal effects is larger. As discussed Section 3, we consider the worst-case change in the distribution of covariates when we estimate external robustness. However, it turns out that these first two studies are in a special scenario where populations differ from the experimental data in the direction toward causal effects of larger magnitude.

## F.3 Incorporating Partial Knowledge about Population Data

Here, we consider the first two studies (Peffley and Hurwitz, 2007; Caprariello and Reis, 2013) and show how we can incorporate knowledge about population data. In particular, we incorporate population proportions of education categories ("Less than College", "College", and "Graduate School"), age categories ("Age between 18 and 39", "Age between 40 and 59", "Age above 60") and white when we estimate external robustness. Then, estimated external robustness improves to 0.75 for Peffley and Hurwitz (2007), and improves to 0.84 for Caprariello and Reis (2013). They now correctly indicate high external robustness as we found in the T-PATE estimates in the national probability surveys.

This additional analysis shows that when partial knowledge of population data is available, it is important to include such information into estimation of external robustness.

# G Simulation Study

In this section, we conduct a simulation study to investigate the finite sample performance of our proposed estimator for external robustness. We demonstrate two key results. First, as expected from our theoretical results, the convergence to the true external robustness is slower if the underlying CATE estimation is more difficult and if the true external robustness is high. Second, the convergence to the true external robustness requires a relatively small sample size, and most importantly, if we consider benchmarks we constructed in Section 4, even with small sample size like 500, estimated external robustness accurately classifies whether the true external robustness is below "small" benchmarks, between "small" and "moderate" benchmarks, or above "moderate" benchmark.

## G.1 Simulation Design

We use $n$ to denote the sample size and we vary $n \in \{500, 1000, 2000, 4000\}$, which covers from small to large experiments. We consider the following data generating process following the design of simulation studies by Wager and Athey (2018). We first draw covariates $\mathbf{X}_i \sim \text{Uniform}([0,1]^K)$ where $K$ is the number of covariates we vary later. Then we set the CATE function. We first define the base CATE function.

$$\widetilde{\tau}(\mathbf{x}) = \{1 + 1/(1 + \exp^{-20(x_1 - 1/3)})\} \times \{1 + 1/(1 + \exp^{-20(x_2 - 1/3)})\}$$

where $x_1$ and $x_2$ are the first two covariates. This CATE function is the same as the one investigated in Wager and Athey (2018). To consider different levels of external robustness, we shift the center of the CATE function. In particular, we define the true CATE function as follows.

$$\tau(\mathbf{X}_i) = \left\{ \widetilde{\tau}(\mathbf{X}_i) - \frac{1}{n} \sum_{j=1}^{n} \widetilde{\tau}(\mathbf{X}_j) \right\} + \nu_i$$

where $\nu_i \sim \text{Normal}(\mu, 0.2)$. Here, $\mu$ is equal to the true SATE, and we vary it later.

Given the true CATE function, we now draw potential outcomes. For potential outcomes under control, $Y_i(0) = -\tau(\mathbf{X}_i)/2 + \epsilon_i$ where $\epsilon_i \sim \text{Normal}(0, 0.5)$. For potential outcomes under control, $Y_i(1) = \tau(\mathbf{X}_i)/2 + \epsilon_i$ where $\epsilon_i \sim \text{Normal}(0, 0.5)$. We consider a randomized experiment where $T_i \sim \text{Bernoulli}(0.5)$. Then, we observe outcomes $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$. For each unit $i$, we observe $(Y_i, T_i, \mathbf{X}_i)$.

From theoretical results in Section 3.3 and Appendix , we know the convergence rate is governed mainly by the two key parameters; (a) how fast the underlying CATE function converges to the true CATE function in $L_2$ (not point-wise), and (b) whether the underlying true external robustness is large. To vary (a), we consider two different numbers of covariates $K \in \{2, 10\}$, which are values investigated also in Wager and Athey (2018). We expect that when $K = 10$, the CATE estimation is more difficult, and the converagence rate of estimated external robustness is slower. To vary (b), we consider three different values for the mean of the CATE

$\mu \in \{0.3, 0.8, 1.3\}$. We choses them such that the true external robustness is about 0.04 (smaller than "small" benchmark of 0.14), 0.28 (between "small" benchmark of 0.14 and "moderate" benchmark of 0.57), and 0.63 (larger than "moderate" benchmark of 0.57). Therefore, in total, we consider 6 different scenarios ($3 \times 2$).

## G.2 Results

Figure 2 shows simulation results. Several points are worth noting. First, in all 6 scenarios, our estimated external robustness converges to the true external robustness as sample size grows. Second, when sample size is smaller (e.g., 500), the bias is larger, and yet, all estimates are below the small benchmark (0.14) in the first row, all estimates are between the small and moderate benchmarks (0.14 and 0.57) in the second row, and all estimates are above the moderate benchmark (0.57) in the third row. Third, as predicted from our theoretical results, when the CATE estimation becomes more difficult (i.e., $K$ is larger; the second column), the convergence becomes slower, while its effect seems to be relatively minor. Finally, as predicted from our theoretical results, when the true external robustness is large (i.e., $\mu$ is larger; the third row), the convergence becomes slower.

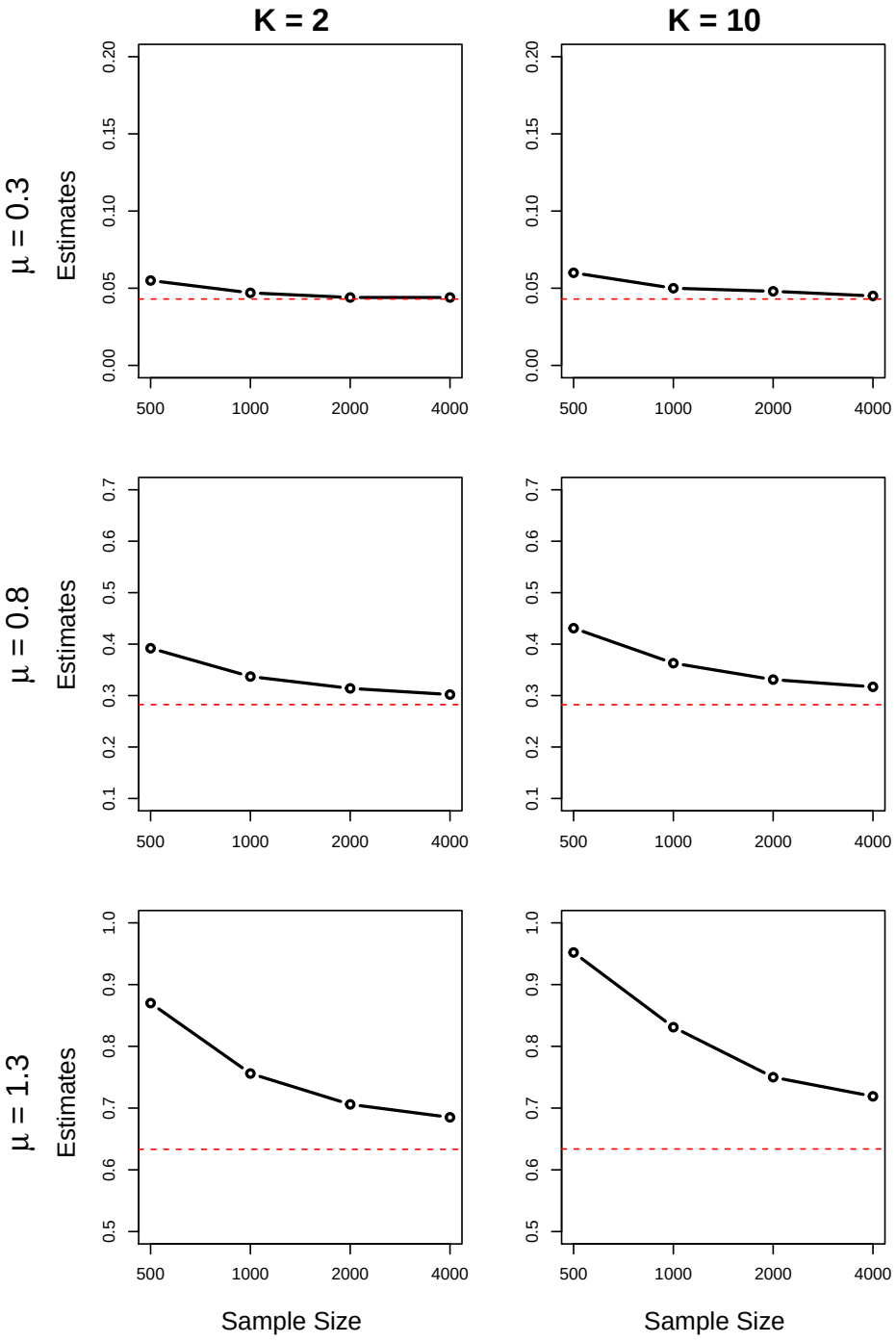



Figure 2: Simulation Results. *Note*: The true external robustness is represented by a red dotted line in each scenario.

# References


Berinsky, Adam J, Gregory A Huber and Gabriel S Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3):351–368.

Boyd, Stephen and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

Coppock, Alexander, Thomas J Leeper and Kevin J Mullinix. 2018. "Generalizability of Heterogeneous Treatment Effect Estimates Across Samples." *Proceedings of the National Academy of Sciences* 115(49):12441–12446.

Domurat, Richard, Isaac Menashe and Wesley Yin. 2021. "The Role of Behavioral Frictions in Health Insurance Marketplace Enrollment and Risk: Evidence From A Field Experiment." *American Economic Review* 111(5):1549–74.

Egami, Naoki and Erin Hartman. 2022. "Elements of External Validity: Framework, Design, and Analysis." *American Political Science Review* .

Johnston, Christopher D and Andrew O Ballard. 2016. "Economists and Public Opinion: Expert Consensus and Economic Policy Judgments." *The Journal of Politics* 78(2):443–456.

Newey, Whitney K and Daniel McFadden. 1994. Large Sample Estimation and Hypothesis. In *Handbook of Econometrics*, ed. Robert Engle and Daniel McFadden. North Holland pp. 2112–2245.

Wager, Stefan and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113(523):1228–1242.