

UC Riverside

UC Riverside Previously Published Works

Title

Identifying Causes of Treatment Effect Heterogeneity Across Replication Studies: Definitions, Identification, and Estimation of Remaining Heterogeneity

Permalink

<https://escholarship.org/uc/item/71s7x1tb>

Journal

Psychological Methods

ISSN

1082-989X

Authors

Park, Soojin

Pohl, Steffi

Steiner, Peter M

Publication Date

2026-09-24

DOI

10.1037/met0000863

Peer reviewed

Identifying Causes of Treatment Effect Heterogeneity Across Replication Studies:
Definitions, Identification, and Estimation of Remaining Heterogeneity

Soojin Park¹

Steffi Pohl²

Peter M. Steiner³

¹School of Education, University of California, Riverside

²Department of Education and Psychology, Freie University

³Department of Human Development and Quantitative Methodology, University of
Maryland

Author Note

©American Psychological Association, 2026. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. The final article is available, upon publication, at: [10.1037/met0000863](https://doi.org/10.1037/met0000863). Research reported in this publication was supported by the National Science Foundation (NSF, 2243119), the Institute of Education Sciences, U.S. Department of Education (R305D220034), and the Deutsche Forschungsgemeinschaft (German Research Foundation, 464369680). The authors acknowledge the assistance of Gemini 2.5 Pro for grammar checking and editing during the manuscript preparation process. Correspondence concerning this article should be addressed to Soojin Park, School of Education, University of California-Riverside, Riverside, CA92521, USA. Email: soojin.park@ucr.edu

Abstract

Replication studies are crucial for evaluating treatment efficacy and exploring the conditions that moderate treatment effects. Understanding this treatment effect heterogeneity across studies is essential for building robust theories. However, identifying the causes of this heterogeneity is challenging, as multiple study characteristics often vary simultaneously, even in replication studies in which many aspects are deliberately held constant across studies. Traditional meta-regression typically lacks a clear causal framework, obscuring the causal interpretation of moderating effects. While a more recent causal method provides a way to identify the replication effect after adjusting for unintended differences, it relies on the strong assumption that all effect moderators that cause the treatment effect heterogeneity across studies must be measured and uses a parametric estimation model. We propose an alternative approach by introducing a causal framework that decomposes sources of treatment effect heterogeneity. Our approach uses a Directed Acyclic Graph (DAG) to identify these sources and to develop strategies to single out their contributions. By focusing on different (causal and non-causal) quantities, the proposed strategies require weaker assumptions than the earlier causal approach. We also propose robust estimation strategies using a variable selection method designed to model these higher-order interactions with minimal bias. We illustrate our method using a real-world example of replication studies from social psychology. Through a simulation and case study, we provide useful recommendations for applied researchers on designing future replication studies and identifying moderators that cause effect heterogeneity.

Identifying Causes of Treatment Effect Heterogeneity Across Replication Studies:
Definitions, Identification, and Estimation of Remaining Heterogeneity

1. Introduction

Replication studies are vital for confirming study findings and clarifying the conditions under which a treatment is effective (Schmidt, 2009; Doabler et al., 2016). *Direct replication* aims to test whether the same conclusions about the effectiveness of a treatment can be obtained using the same methods and procedures. In contrast, *conceptual replication* systematically varies a certain study characteristic by design (while keeping others constant) to evaluate the generalizability of the effect and to identify conditions that potentially moderate the treatment effect. While treatment effect heterogeneity is an expected outcome of conceptual replications, it also frequently emerges in direct replications due to inevitable differences in context, such as study populations or settings, across studies (McShane, Tackett, Böckenholt, & Gelman, 2019; Nosek & Errington, 2020). Understanding this heterogeneity is essential as it strengthens theory by specifying the scope of an intervention and the conditions under which a treatment is effective (Linden & Hönekopp, 2021).

However, identifying causes of treatment effect heterogeneity is challenging, even in replication studies where potential moderators are carefully manipulated by design. This is because many study characteristics often vary simultaneously across studies. While varying only one aspect at a time and keeping all other aspects constant across studies is ideal, it is often impossible in practice. For example, consider two replication studies of an online intervention to reduce prejudices that yield differences in treatment effects (e.g., Twardawski et al., 2026). Deliberately varying the recruitment strategy by design, say, from contacting participants based on a university-hosted email list to contacting them via an online platform, does not only change participant characteristics, such as age and education, but also setting characteristics, such as the device used for the survey (phone, tablet, computer) or the noise level participants experience. Is the treatment effect

heterogeneity due to the shifting participant characteristics or the setting in which the intervention was delivered? Such simultaneous variations in population and setting characteristics makes it difficult to isolate and assess the impact of specific causes on observed effect heterogeneity.

Meta-analytic models are effective for identifying an average treatment effect across studies and describing correlates of effect heterogeneity. They do, however, usually not aim to identify the causes of effect heterogeneity. As such, these models often lack formally defined causal estimands (effects of interest), explicit identification assumptions, and estimation targeted at causal effects. Also, individual studies are rarely designed for cross-study comparisons of causal claims. They often lack the relevant variables that are necessary to satisfy the conditional independence assumption. Even when relevant variables are measured, they are not assessed consistently across studies. This precludes researchers from accounting for them effectively. To draw a causal interpretation regarding the impact of intervention type (e.g., variation in treatment format or delivery) on the treatment effect, one needs to 1) account for the causal ordering among the typically correlated moderators and 2) clarify any potential moderator-outcome confounding.

To address this issue, Pohl et al. (in press) proposed a causal framework to identify the remaining heterogeneity after adjusting for unintended differences in replication studies. While this approach allows for identifying specific study characteristics (such as population or setting) as causes of treatment effect heterogeneity, it relies on the strong assumption that all moderators that explain the effect heterogeneity between the studies are measured. Furthermore, it relies on parametric estimation methods, which are prone to model misspecification. This study differs from this prior work in two critical dimensions. First, we shift the focus from the causal effect of study characteristics to the decomposition of observed effect heterogeneity between replication studies. By shifting the focus to these decomposed quantities, some causal and others non-causal, our framework requires weaker assumptions. Second, we introduce a flexible estimation framework, which is able to

accommodate high-dimensional interactions and nonlinearities to minimize the risk of model misspecification.

Specifically, when identifying moderators, we first develop a causal Directed Acyclic Graph (DAG, Pearl, 2009) of the presumed data-generating process of the replication studies (Section 2). Guided by DAG, we focus on 1) estimating the overall treatment effect heterogeneity across studies, 2) explaining treatment effect differences across studies using baseline moderators and 3) identify a single causal moderator of treatment effects (Section 3). In particular, we delineate which kind of effect moderators can be identified as causing effect heterogeneity and which can only be used for describing effect heterogeneity. By differentiating between these two kinds of sources, our method benefits from relatively weaker assumptions.

In Section 4, we propose two effect-heterogeneity estimators: a coefficient-based estimator and an imputation estimator. For the imputation estimator, we discuss a flexible modeling strategy based on a variable selection method to capture higher-order interactions. We also briefly demonstrate the performance of the estimators in a simulation study.

Finally, we demonstrate the usefulness of our proposed methodology using a case study (Section 5) and we conclude with a discussion (Section 6). The step-by-step implementation guide in R appears in Online Supplementary Material E. The R code to replicate the simulation and case studies is available at the first author’s GitHub repository (Park, 2026): <https://github.com/soojinpark33/replication>.

2. Running Example

To motivate the proposed methodology, we consider a systematically planned replication study from Twardawski et al. (2026) that involves two online randomized experiments with different recruitment strategies. The first study ($R = 0$) used a convenience sample of unpaid participants recruited through university-hosted e-mail lists

and advertisements (hereafter, referred as “unpaid university sample”). The second study ($R = 1$) replicated the same experiment but used a sample of paid participants recruited from the academic online platform **Prolific** (hereafter, referred as “paid provider sample”). The experiment examined the effect of the imagined intergroup contact (X) on attitudes towards homeless persons (Y). Participants of each study were randomly assigned to either a treatment condition that required the imagination of a positive interaction with a homeless person ($X = 1$) or a control condition that required the imagination of walking in a forest ($X = 0$). The outcome (Y) was the participants’ attitude toward homeless persons, measured after the imagination task.

While the two studies used the exact same treatments, outcome measures, and experimental design, the deliberate variation in recruitment strategy, which is our design difference, led to differences in several study characteristics that could potentially moderate the treatment effects. These include person characteristics Z (e.g., age, education) as well as setting characteristics W during the experiment (e.g., device type, location). Specifically, the paid provider sample includes older participants than the unpaid university sample. Then, the difference in the age distribution likely impacted also setting variables, as younger participants were more likely to use a mobile phone for the study, while older participants were more likely to use a laptop at home. The differences captured by replication factor R reflect these factors that vary between the two samples due to systemic differences in recruitment strategies. The list of all measured variables of person characteristics and setting characteristics is provided in Table S1 of Online Supplementary Material A. Importantly, these characteristics have the potential to moderate the effect of the treatment on the outcome.

Thus, unsurprisingly, Twardawski et al. (2026) found that the average treatment effect (ATE) was larger in the paid provider sample than in the unpaid university sample. This finding raises an important question: what are (some of) the causes of this observed effect heterogeneity? Researchers might be interested in (1) investigating the extent of

effect heterogeneity that is (causally) attributable to person or setting characteristics of interest and (2) obtaining an estimate of the effect heterogeneity remaining after the removal of any effect differences due to specific person or setting characteristics.

Causal DAG of the Replication Study. To define attributable and remaining effect heterogeneity, we formalize the presumed data-generating process of the two replication studies in a stylized causal DAG (see Figure 1). Causal DAGs are formal objects used to encode a researcher’s qualitative background knowledge and assumptions about the real-world process that generated the observed data. That is, they represent the (nonparametric) structure of causal relationships among both observed and unobserved variables involved in the data-generating process (DGP). A DAG is causal if it includes all common causes of each pair of variables in the graph, implying that observed associations among variables can only be transmitted along causal and noncausal paths represented in the DAG. Importantly, together with structural causal models, causal DAGs formally link qualitative knowledge about the DGP to probability theory and statistics.

The strength of causal DAGs lies in their use. They can be used to define causal effects, assess whether causal effects are identifiable from the observed data, select covariate sets that remove confounding bias, assess the direction and magnitude of any remaining bias, and clearly communicate the key assumptions on which causal analyses rest (Quimpo & Steiner, 2026). For a given causal DAG, identification analysis and covariate selection can be fully automated and implemented, for instance, in DAGitty (Textor, van der Zander, Gilthorpe, Liśkiewicz, & Ellison, 2016). A “stylized” causal DAG is a graph used to convey the main ideas for methodological purposes but does not fully explicate all variables involved in the DGP (and therefore cannot be directly used in practice for formal, automated identification analyses). For an introduction to causal DAGs, see Feng (2024), Pearl (2010), or Steiner, Shadish, and Sullivan (2023). For guidance on drawing credible DAGs in applied research, see Quimpo and Steiner (2026).

We now construct a stylized causal DAG for our empirical example. Within each

study, we are interested in the causal impact of treatment X (the imagined intergroup contact) on the outcome Y (the attitude towards homeless persons). Due to random assignment, X and Y are only connected by the causal path $X \rightarrow Y$ but no other, noncausal path. Thus, due to the absence of any confounding, the causal effect is directly identified without the need for covariate adjustment.

Because the data are from two studies, the outcome may also directly depend on the replication factor R , as shown by the causal link $R \rightarrow Y$. Since the outcome Y is causally determined by treatment exposure X and replication factor R , they may interact in producing the outcome ($R \rightarrow Y \leftarrow X$), that is, study differences may moderate the causal effect of X on Y .¹ This type of moderation effect is the treatment effect heterogeneity we are interested in. To highlight that we are interested in the causal effect of X on Y *moderated* by replication status R , we graphically represent it as $R \rightarrow \{Y \leftarrow X\}$.

Since the DAG also contains mediators W and W' , the moderated effect along $R \rightarrow \{Y \leftarrow X\}$ represents only the direct effect. The indirect effects are those mediated through W and W' , capturing how the replication status R indirectly impacts the outcome Y and how the mediators contribute to the effect heterogeneity $R \rightarrow \{W, W'\} \rightarrow \{Y \leftarrow X\}$. We define two sets of mediators: W' , the single or multiple target mediator(s), and W , a set of other mediators that may causally affect the target mediator W' . In our example, W' represents the participants' device choice and W represents location, noise level, temperature, and presence of others during the experiment. Note that only mediators that can potentially vary with R are considered here. This is, if the participants in one study (the unpaid university sample) would have been recruited into the other (the paid provider sample), their mediator values could have been different (at least for some participants).

The causal graph further shows that R 's causal impact on Y (including its moderating impact on $Y \leftarrow X$) is confounded by observed baseline covariates Z , as shown

¹ Note that causal DAGs are nonparametric representations of the data-generating process and, thus, automatically include all possible interaction patterns of a variable's causes.

by the confounding path $R \leftarrow Z \rightarrow \{Y \leftarrow X\}$. In our example, all person characteristics (age, gender, language, social status, prior knowledge in social sciences, psychological characteristics such as motivation) may simultaneously affect the likelihood of being in the convenience or provider sample (R) and the outcome Y . These baseline covariates may also confound the causal relations between R and $\{W, W'\}$, between W and W' , and between $\{W, W'\}$ and Y and, thus, further contribute to the effect heterogeneity across studies.

Finally, it is also possible that unobserved variables U confound the causal relation between all observed variables. The dashed arrows indicate that, for some of the proposed analyses, we must later assume that U has no impact on the outcome Y and the target mediator W' . This assumption implies that all relevant confounders of the W' – Y relationship are fully captured by the observed variables R, Z and W . In our case study, the time of the day the study was conducted may affect setting variables (such as noise level or temperature), but may not be relevant for prejudices (Y). Such variables are subsumed in U in the DAG.

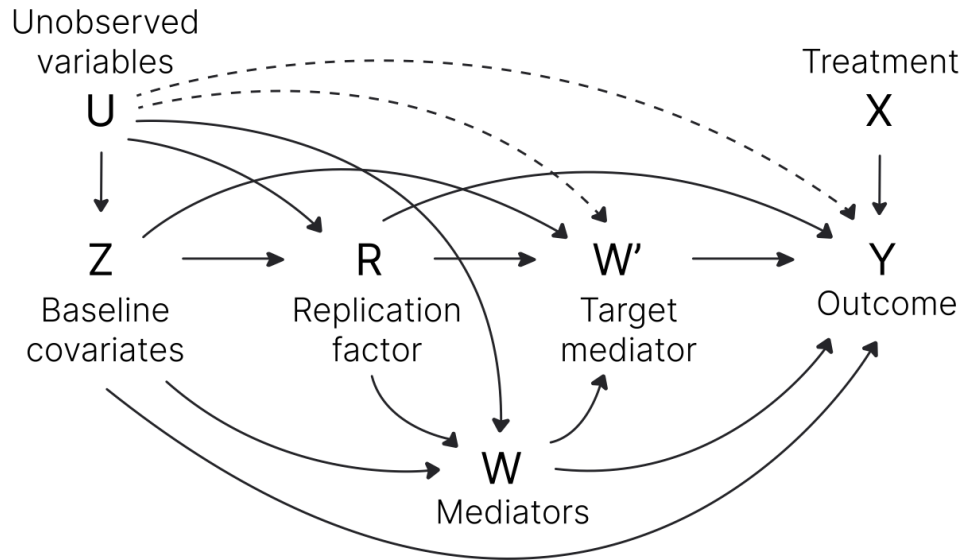


Figure 1. Causal DAG of the Replication Study

Note: The dashed arrows indicate that, for some analyses, we need to assume that unobserved variables U do not affect the outcome and the target mediator.

Inspecting all paths connecting R and Y in the causal DAG reveals the sources of

observed effect heterogeneity across the two studies. The heterogeneity in the causal effect $X \rightarrow Y$ is transmitted along three different types of paths:²

P1. Non-causal paths from replication status R to the outcome Y through observed person characteristics Z or unobserved baseline variables U (overall, there are 12 open non-causal paths; e.g., $R \leftarrow Z \rightarrow \{Y \leftarrow X\}$ or $R \leftarrow U \rightarrow Z \rightarrow W \rightarrow \{Y \leftarrow X\}$).

P2. Causal paths from R to Y through the target mediator W' (two paths:

$$R \rightarrow W' \rightarrow \{Y \leftarrow X\} \text{ and } R \rightarrow W \rightarrow W' \rightarrow \{Y \leftarrow X\}).$$

P3. Causal paths from R to Y *not* traversing through W' (two paths: $R \rightarrow \{Y \leftarrow X\}$ and $R \rightarrow W \rightarrow \{Y \leftarrow X\}$).

It is important to note the differences between these paths. The P1 paths operate through baseline characteristics Z and U that are confounders (common causes) on these paths. Thus, these paths are *non-causal* and the effect-moderating associations transmitted along them cannot be causally attributed to R . Furthermore, interpreting the causal effect of these baseline covariates Z is difficult, as it requires conceptualizing interventions on socially ascribed characteristics like age or gender (see e.g., VanderWeele & Robinson, 2014; Lundberg, 2020). Given these challenges with baseline variables, we abstain from a *causal* interpretation and instead use a descriptive, associational interpretation instead. That is, we will examine effect heterogeneity within the same levels of Z and then evaluate Z 's contribution to the overall effect heterogeneity when adjusting for Z .

In contrast, the P2 and P3 paths are causal paths, as they represent the direct or indirect effects of being in a different replication study. For paths P2, we can formally define hypothetical interventions on the target mediator W' and causally interpret their effect-moderating impact on the treatment effect (assuming no unobserved confounding). In our example, the causal paths P2 may arise if specific platforms (e.g., Prolific) impose

² We do not discuss any naturally closed collider paths because they do not transmit any effect-moderating associations.

constraints that directly influence device choice. Paths P3 arise if participating via a specific platform affects the location and noise level from which one participates (W), which in turn influences the outcome and moderates the treatment effect.

Guiding Questions. Given the different sources of effect heterogeneity, we will address three specific questions:

1. What is the magnitude of the overall treatment effect heterogeneity between the two studies—that is, the heterogeneity induced along all three path types (P1, P2, and P3)?
2. How much of this overall heterogeneity is explained by differences in baseline covariates Z between the studies—that is, the heterogeneity induced by paths through Z in P1?
3. How much of the remaining heterogeneity, after adjusting for baseline differences, is *causally* attributable to the target mediator W' —that is, the heterogeneity induced along paths P2?³

3. Definitions and Identification Results

In this section, we introduce our framework of decomposing the observed treatment effect heterogeneity across the two studies (i.e., their difference in the effect estimates). Specifically, we isolate the portion of the effect heterogeneity explained by differences in baseline covariates, the portion causally attributable to the target mediator, and the remaining heterogeneity left after accounting for these factors. We begin by defining the causal effect of each individual study and the treatment effect heterogeneity across them. In this section, we focus on identification: the process of determining when inherently

³ Researchers may also be interested in the effect heterogeneity induced by all associations between R and Y that are transmitted via W' , including those via (specific) variables in Z , rather than only the heterogeneity caused directly by R . While this represents a significant question, it remains beyond the scope of this study. Such inquiry requires different estimands and assumptions, which we leave for future research.

unobservable counterfactual quantities can be expressed in terms of potentially observable data.

3.1. Causal Effects of Individual Studies

For each of the two replication studies, $R \in \{0, 1\}$, we are interested in the average treatment effect (ATE) of a well-defined treatment-control contrast $X \in \{0, 1\}$ on an outcome measure Y for the study-specific population and setting. Using potential outcomes notation (Imbens & Rubin, 2015), the study-specific ATEs are defined as

$$ATE_r = E[Y(1) - Y(0)|R = r]. \quad (1)$$

For each study, the ATE is identified if the following assumptions hold:

A1. *Independence (randomization of treatments):* $Y(x) \perp X|R = r$ for any $x, r \in \{0, 1\}$.

Since we assume that our studies are RCTs, no conditioning on other covariates is needed. For non-randomized studies one can formulate a corresponding conditional independence assumption.

A2. *Positivity:* $0 < P(X = x|R = r, Z = z)$ for any $x, r \in \{0, 1\}$ and $z \in \mathcal{Z}$ (the set of all observed covariate values z). This is, within each study, the conditional probability of receiving any treatment condition must be nonzero. We consider the probability conditional on Z because it is needed for the identification results across studies we discuss below.

A3. *Consistency and no interference:* $Y = XY(1) + (1 - X)Y(0)$ for any $r \in \{0, 1\}$. For each study, the observed outcome must be the same as the potential outcome under the actually realized treatment condition X . This implies that there are no hidden variations in the treatment and control conditions. Moreover, the observed outcome only depends on the participants own treatment status but not on other participants' treatment exposure (no interference).

In our example, assumptions A1 and A2 are met if the RCTs were perfectly implemented (including no attrition and nonresponse). Assumption A3 is met unless participants would have received different versions of treatment conditions or their outcomes are affected by another participants' treatment status.

Under assumptions A1-A3, the study-specific ATE_r , for $r \in \{0, 1\}$, is nonparametrically identified given the observed data (Imbens & Rubin, 2015):

$$ATE_r = E[Y(1) - Y(0)|R = r] = E[Y|X = 1, R = r] - E[Y|X = 0, R = r]. \quad (2)$$

That is, the simple difference in the expected outcomes of the treatment and control group identifies the study-specific ATE.

3.2. First Question: Treatment Effect Heterogeneity

To address the first guiding question, we define the treatment effect heterogeneity (τ) as the difference between the two study-specific ATEs:

$$\tau = ATE_1 - ATE_0 = E[Y(1) - Y(0)|R = 1] - E[Y(1) - Y(0)|R = 0]. \quad (3)$$

Differences in ATEs may arise from variations in units (i.e., participants), treatments, outcomes, or settings across studies (UTOS; Cronbach, 1982). However, for systematically planned replication studies as in our example, we can reasonably rule out some of these sources by design. Specifically, the treatment conditions and outcome measures were identical across studies (i.e., no hidden variation in treatments across studies). Further, the design of the online experiments made any interference effects (spill-over or treatment contamination) unlikely, both within and across studies. By ruling out these sources of effect heterogeneity, any remaining effect heterogeneity can then be attributed to differences in person or setting characteristics.

Regardless of its source, ATE heterogeneity is directly identified as the difference

between the study-specific ATE estimates:

$$\begin{aligned}\tau &= ATE_1 - ATE_0 \\ &= (E[Y|X = 1, R = 1] - E[Y|X = 0, R = 1]) - \\ &\quad (E[Y|X = 1, R = 0] - E[Y|X = 0, R = 0]).\end{aligned}\tag{4}$$

Conceptually, τ quantifies the degree of treatment effect heterogeneity across replication studies, as outlined in the first guiding question. Since the effect heterogeneity τ is a descriptive measure with regard to two causal estimands, no further assumptions other than A1-A3 are needed. However, as discussed above, if certain sources of effect heterogeneity can be reasonably ruled out, a more meaningful interpretation of any observed heterogeneity is possible. With carefully planned systematic replication studies, researchers might be able to attribute the ATE differences to a single or a few specific causes (Twardawski et al., 2026; Steiner, Wong, & Anglin, 2019; Wong, Anglin, & Steiner, 2022).

3.3. Second Question: Effect Heterogeneity Explained by Baseline Covariates Z

To address the second guiding question, we decompose the total effect heterogeneity into the portion explained by the differences in baseline covariates Z and the remaining unexplained heterogeneity. The remaining heterogeneity (ζ_Z) quantifies the average difference in treatment effects between the two studies after adjusting for baseline covariates $Z = z$. That is, ζ_Z represents the portion of the average difference in treatment effects that cannot be accounted for by differences in baseline covariates Z (person characteristics in our example):

$$\zeta_Z = \sum_Z \left\{ \underbrace{E[Y(1) - Y(0)|R = 1, Z = z]}_{\text{ATE in Replication Study at } Z = z} - \underbrace{E[Y(1) - Y(0)|R = 0, Z = z]}_{\text{ATE in Original Study at } Z = z} \right\} \underbrace{P(Z = z)}_{\text{Weighting by } Z} \tag{5}$$

Then, the effect heterogeneity explained by baseline covariates Z is given by the difference between the unadjusted (τ) and remaining effect heterogeneity: $\delta_Z = \tau - \zeta_Z$. In our

example, the difference in the two heterogeneity measures describes the amount of heterogeneity explained by differences in the studies' person characteristics. Note that δ_Z does not warrant a causal interpretation because Z does not need to be a cause of Y and, moreover, Z is very likely associated with other (causal) effect moderators (for instance, with U in the causal DAG in Figure 1).

If the causal assumptions A1-A3 hold for each study, we can identify the remaining heterogeneity, ζ_Z , and the heterogeneity explained by baseline covariates, δ_Z , because the term $E[Y(1) - Y(0)|R = r, Z = z]$ in equation (5) is nonparametrically identified as:

$$E[Y(1) - Y(0)|R = r, Z = z] = E[Y|R = r, X = 1, Z = z] - E[Y|R = r, X = 0, Z = z], \quad (6)$$

for $r \in \{0, 1\}$. Using this identification result, the potential-outcomes-based heterogeneity definition in equation (5) can be expressed in terms of observed data.

3.4. Third Question: Effect Heterogeneity Causally Attributable to W'

To address the third guiding question, we further decompose the remaining heterogeneity by evaluating a hypothetical intervention on the target mediator W' . This allows us to identify *causal* moderators of treatment effects. Specifically, we evaluate two hypothetical interventions, conditional on Z : 1) **static intervention**: fixing the target mediator at a specific value $W' = w'$, and 2) **stochastic intervention**: equalizing the distribution of the target mediator W' across studies. The choice between these interventions depends on the substantive question. In our example, the static case involves a scenario where all participants, regardless of their study group, conduct the experiment on a laptop. This hypothetical intervention effectively removes device-related variations, simulating the conditions of a perfectly controlled experimental setting. The stochastic case considers a scenario in which the paid provider sample has the same distribution of device types (e.g., same proportion of laptop users) as the unpaid university sample. This approach simulates a scenario where a strict controlled environment is implausible, but

achieving a similar distribution in the target mediator is still desired.

It is important to note that the following estimands and estimators of remaining heterogeneity are marginalized over the combined distribution of baseline covariates (Z) from both studies, $P(Z)$. However, depending on the research question, researchers may choose to marginalize with respect to any other reference population. For example, they could use the observed covariate distribution of one of the two studies (i.e., $P(Z|R = r)$ for $r \in \{0, 1\}$).

Scenario 1: Fixing the Target Factor. Upon observing remaining heterogeneity even after adjusting for baseline covariates, researchers may ask whether the heterogeneity would persist if both studies were conducted under identical conditions—for instance, with a standard laptop. The remaining heterogeneity at a specific level of the target moderator ($\zeta_{w'}$) captures the expected difference in treatment effects between the two studies when the target mediator is fixed at $W' = w'$, after adjusting for baseline covariates Z :

$$\zeta_{w'} = \sum_Z \left\{ \underbrace{E[Y(1, w') - Y(0, w')|R = 1, Z = z]}_{\text{ATE in Replication after fixing } W'=w'} - \underbrace{E[Y(1, w') - Y(0, w')|R = 0, Z = z]}_{\text{ATE in Original at after fixing } W'=w'} \right\} \underbrace{P(Z = z)}_{\text{Weighting by } Z} \quad (7)$$

for $w' \in \mathcal{W}'$, where $Y(1, w')$ is the potential outcome when the individual received treatment $X = 1$ in setting $W' = w'$. Note, this definition is based on two interventions, on the treatment X and the target moderator W' . Therefore, the potential outcomes depend on both interventions. As a result, the remaining heterogeneity ($\zeta_{w'}$) captures the expected ATE difference between the two studies in setting $W' = w'$, after the removal of any effect heterogeneity explained by baseline covariates Z .

The effect heterogeneity causally attributable to fixing the target moderator $W' = w'$, after controlling for baseline covariates, is then given by the difference in remaining heterogeneity before and after fixing $W' = w'$: $\delta_{w'} = \zeta_Z - \zeta_{w'}$.

In our example, remaining heterogeneity after this static intervention ($\zeta_{w'}$) represents the ATE difference that were observed if all participants had conducted the experiment on

a laptop, after adjusting for person characteristics. Consequently, the difference in remaining heterogeneity before (τ_Z) and after ($\zeta_{w'}$) fixing $W' = w'$ represents the effect heterogeneity that is attributable to this static intervention.

As the definition involves potential outcomes $Y(x, w')$, the data-generating process (as outlined in the causal graph) must meet the following assumptions in addition to study-specific assumptions A1–A2 (A3 is covered by B3 below):

B1. *No unmeasured W' -specific effect moderators:*

$W' \perp \{Y(1, w') - Y(0, w') | R = r, Z = z, W = w \text{ for any } r \in \{0, 1\}, z \in \mathcal{Z} \text{ and } w \in \mathcal{W}\}$. That is, there are no unmeasured treatment effect-moderators that influence the target factor W' , conditional on R, Z , and W .

B2. *Positivity regarding the target factor W' :*

$0 < P(W' = w' | R = r, X = x, Z = z, W = w)$. That is, within each study, the probability of receiving any level of W' conditional on X, Z , and W must be nonzero.

B3. *Consistency and no interference:* $E[Y(x, w') | R = r, X = x, Z = z, W = w, W' = w'] = E[Y | R = r, X = x, Z = z, W = w, W' = w']$. For each study, and conditional on Z and W , the average observed outcome must be the same as the average potential outcome after intervening to set treatment X and the target factor W' to (x, w') . The no interference implies that a participant's treatment and setting exposure (x, w') has no impact on other participants' potential outcomes (within and across the two studies).

All three assumptions are strong and must be carefully assessed in practice. Table 1 provides further details on their implications and plausibility in our example. In many applied settings, these assumptions may not be fully satisfied. For example, the assumption of no unmeasured W' -specific effect moderators (B1) is violated if a key baseline confounder, such as participant age, is omitted from the data. Furthermore, even when all confounders are measured, residual confounding still remains when those variables are

measured with error. In such cases, a strict causal interpretation may be fragile or not warranted. Nonetheless, we argue that the resulting estimates are closer to the true causal effect than the standard meta-analysis associations, as we considered causal ordering between confounders and carefully conditioned a set of confounders based on a DAG. Furthermore, research has shown that the biasing impact of unobserved variables and measurement error in covariates may be partly compensated by other observed covariates (Cook, Steiner, & Pohl, 2009; Sengewald, Steiner, & Pohl, 2019; Steiner, Cook, & Shadish, 2011; Steiner & Kim, 2016). The bias introduced by measurement error in covariates may even be fully addressed through the use of psychometric models (Lockwood & McCaffrey, 2014, 2015, 2016; Sengewald & Pohl, 2019).

Under assumptions A1–A2 and B1–B3, the remaining heterogeneity $\zeta_{w'}$ is nonparametrically identified because the difference in expected potential outcomes is recoverable from observed conditional expectations and probability distributions:

$$\begin{aligned}
 & E[Y(1, w')|R = r, Z = z] - E[Y(0, w')|R = r, Z = z] \\
 &= \sum_w \{E[Y|R = r, X = 1, Z = z, W = w, W' = w'] \\
 &\quad - E[Y|R = r, X = 0, Z = z, W = w, W' = w']\}P(W = w|R = r, Z = z),
 \end{aligned} \tag{8}$$

for each study $R = r$. Thus, both the remaining and attributable effect heterogeneity are identified given the observed data. A proof for this identification result is shown in Online Supplementary Materials B.

Scenario 2: Equalize the Target Distribution Across Studies. While fixing both studies to the same setting ($W' = w'$) often addresses a meaningful research question about the effect heterogeneity across two studies, the attributable and remaining heterogeneity generally depends on the chosen value of w' . However, researcher may not directly be interested in the effect heterogeneity if all participant in our example would have used the same device, but instead be interested in the change in effect heterogeneity if

Table 1

Key Assumptions and Their Application to the Example

Assumption	Example
B1. <i>No unmeasured W'-specific effect-moderators</i> Requires that, conditional on observed covariates, there are no unmeasured effect-moderators affecting the target mediator W'. Given that identifying such moderators directly can be empirically challenging, a practical approach is to identify and control for confounders of the $W' - Y$ relationship that could potentially moderate the treatment effect.	This assumption is not empirically testable and must be justified using subject-matter knowledge and causal reasoning (e.g., by using causal DAGs). Any variables that affect both device choice (W') and prejudices (Y) must be considered and included in the set of baseline covariates (Z) or other mediators (W). For example, age (Z) may be a confounder as younger individuals may have less prejudices towards foreign persons and are more likely use mobile phones instead of laptops. Similarly, location (W) during the study may impact both device choice and prejudices. Participating while away from home increases the likelihood of using a mobile phone and may introduce more distractions that diminish the intervention's impact. In this example, age and location need to be considered in the analyses to fulfill this assumption.
B2. <i>Positivity</i> Requires sufficient overlap in covariate distributions such that all levels of the target mediator are observed across covariate strata. Violations require extrapolation, which may not be valid.	Within each study, intervention status, and confounders (e.g., age and location), there must be individuals who both use and do not use laptops. This assumption is violated if, for instance, laptops are only used in one study, only by older individuals, or only within the control group. Researchers can empirically assess this by examining covariate overlap across levels of W'. When overlap is limited, bias can be reduced using trimming or truncation methods (Imbens & Rubin, 2015).
B3. <i>Consistency and No Interference</i> Consistency requires that the treatment and mediator correspond to well-defined interventions. No interference requires that one individual's treatment status does not affected the outcomes of others.	Consistency is fulfilled when the intervention (X) and the devices (W') are similar across participants. This assumption would be violated if some participants received extra intervention sessions or if laptop users utilize devices with dramatically different screen sizes. No interference is fulfilled when the prejudice (Y) of a participant is independent from the intervention status (X) and the choice of device (W') of other participants. The assumption would be violated if participants interact, for example, when a participants that is assigned to the intervention group, shares its insights with other participants. It is also violated when a friend participating nearby on a mobile phone creates a distraction. When there is no interaction between participants within or across the two studies, no interference is likely to be satisfied.

the device distribution in the second (replication) study would have been identical to the observed distribution in the first study. In this case, we can use a stochastic intervention that equalizes the distribution of W' across the two studies rather than fixing it to a single value.

Thus, the research questions here is about how much of the treatment effect heterogeneity persists if we set the distribution of the target moderator W' in the replication study to match the observed distribution of the first study within the same levels of baseline covariates? Formally, the remaining heterogeneity ($\zeta_{W'}$) after this stochastic intervention is defined as:

$$\zeta_{G_{W'}} = \sum_Z \left\{ \underbrace{E[Y(1, G_{W'|R=0,z}) - Y(0, G_{W'|R=0,z}) | R=1, Z=z]}_{\text{ATE in Replication under Original } W' \text{ Distribution}} - \underbrace{E[Y(1) - Y(0) | R=0, Z=z]}_{\text{Observed ATE in Original Study}} \right\} \underbrace{P(Z=z)}_{\text{Weighting by } Z} \quad (9)$$

where $G_{W'|R=0,z}$ denotes, for each participant in the second study, a random draw from the W' distribution of the original study's ($R=0$) within the same levels of baseline covariate ($Z=z$). The term $\zeta_{W'}$ is the remaining ATE heterogeneity that would be observed after equating the distribution of W' within the same levels of baseline covariates.

The effect heterogeneity causally attributable to this stochastic intervention is then given by the difference in remaining heterogeneity before and after equalizing the distribution of W' : $\delta_{G_{W'}} = \zeta_Z - \zeta_{G_{W'}}$.

Under assumptions A1-A2 and B1-B3, the expected difference in potential outcomes

is nonparametrically identified:

$$\begin{aligned}
& E[Y(1, G_{W'|R=0,z}) - Y(0, G_{W'|R=0,z}) | R = 1, Z = z] \\
&= \sum_{w,w'} \{E[Y | R = 1, X = 1, Z = z, W = w, W' = w'] \\
&\quad - E[Y | R = 1, X = 0, Z = z, W = w, W' = w']\} \\
&\quad \cdot P(W = w | R = 1, Z = z) P(W' = w' | R = 0, Z = z).
\end{aligned} \tag{10}$$

Thus, the remaining and attributable heterogeneity are also identified. A proof of this identification result is shown in Online Supplementary Materials C.

4. Estimation

Given the identification assumptions hold, these nonparametric identification results in Sections 3.2 and 3.3 can be used to estimate the target estimands. However, purely nonparametric estimation is often infeasible with finite data, especially when observations are sparse across covariate strata. Therefore, we use parametric models in this section to achieve more efficient and stable estimates of the effect heterogeneity. Note that even if the identification assumptions hold, unbiased effect estimates additionally require the modeling assumptions made in the estimation to hold.

4.1. Coefficient-Based Estimation in Linear Models

Before turning to our general estimation method, we demonstrate how the nonparametric identification results can be expressed in terms of analytic linear models. Assume that the target factor W' and the outcome Y are all continuous. Consider the following two linear models for the outcome Y :

$$\begin{aligned}
& E[Y | R, X] = \tau + \tau_r R + \tau_x X + \tau_{rx} RX, \text{ and} \\
& E[Y | R, X, Z] = \phi + \phi_r R + \phi_x X + \phi_z Z + \phi_{rx} RX + \phi_{xz} XZ + \phi_{rz} RZ + \phi_{rxz} RXZ,
\end{aligned} \tag{11}$$

In these models, we allow treatment effects to vary across replication studies and covariate strata, as represented by the interaction term RX and RXZ , respectively. The first model allows us to estimate the overall treatment effect heterogeneity (τ) via $\hat{\tau}_{rx}$. From the second model, we obtain an estimator for the remaining heterogeneity after adjusting for Z (ζ_Z): $\hat{\zeta}_Z = \hat{\phi}_{rx} + \hat{\phi}_{rxz}\hat{E}[Z]$. The heterogeneity attributable to Z is then obtained as $\hat{\delta}_Z = \hat{\tau}_{rx} - \hat{\zeta}_Z$.

To estimate the remaining heterogeneity after applying static and stochastic interventions and adjusting for Z , we additionally consider the following linear models:

$$\begin{aligned} E[W'|R, Z] &= \beta + \beta_r R + \beta_z Z + \beta_{rz} RZ, \text{ and} \\ E[Y|R, X, Z, W, W'] &= \gamma + \gamma_r R + \gamma_x X + \gamma_z Z + \gamma_w W + \gamma_{w'} W' + \gamma_{xw'} XW' \\ &\quad + \gamma_{rx} RX + \gamma_{rxw'} RXW' + \gamma_{xz} XZ + \gamma_{rz} RZ + \gamma_{rxz} RXZ. \end{aligned} \quad (12)$$

In these models, we allow the treatment effect to vary with the target factor W' , replication status R , and baseline covariates Z (as reflected by the interaction terms XW' , RX , RXW' , XZ , RZ , and RXZ). Based on these models, the remaining heterogeneity after fixing $W = w'$ can be estimated as $\hat{\zeta}_{w'} = \hat{\gamma}_{rx} + \hat{\gamma}_{rxw'}w' + \hat{\gamma}_{rxz}\hat{E}[Z]$, and the heterogeneity attributable to this intervention as $\hat{\delta}_{w'} = \hat{\phi}_{rx} - \hat{\zeta}_{w'}$. This result implies that 1) the remaining heterogeneity depends on the chosen value of w' and 2) no heterogeneity remains if the treatment effects are constant across studies (i.e., $\hat{\gamma}_{rx} = \hat{\gamma}_{rxw'} = \hat{\gamma}_{rxz} = 0$).

Additionally, the remaining heterogeneity after the intervention of equalizing the distribution of the target factor W' is $\hat{\zeta}_{G_{W'}} = \hat{\gamma}_{rx} + \hat{\gamma}_{rxw'} \sum_z E[W'|R=0, z]P(z) + \hat{\gamma}_{rxz}\hat{E}[Z]$. The heterogeneity attributable to this stochastic intervention is captured by $\hat{\delta}_{G_{W'}} = \hat{\phi}_{rx} - \hat{\zeta}_{G_{W'}}$, which can also be written as $\hat{\delta}_{G_{W'}} = (\hat{\beta}_r + \hat{\beta}_{rz}E[Z]) \times (\hat{\gamma}_{xw'} + \hat{\gamma}_{rxw'})$. These expressions imply that $\delta_{G_{W'}}$ would be zero if 1) the treatment effects are constant across studies, baseline covariates, and target factor values, or 2) the distribution of W' does not vary between studies given baseline covariates. Proofs for these parametric identification results are provided in Online Supplementary Materials D.

This coefficient-based approach yields valid estimates for the proposed effects as long as the linear modeling assumptions shown in equations (11) and (12) hold. However, the expressions must be re-derived every time the model changes and are limited to continuous outcomes. To overcome this challenge, we also propose a more flexible general approach in Section 4.2.

Comparison to Meta-regression. Traditional meta-regression via study random-effects models (provided one has a sufficiently larger number of replication studies) or corresponding fixed-effects models can, under specific assumptions, approximate the effect heterogeneity defined in our framework. For example, if we assume the baseline covariates Z are mean-centered such that $E[Z] = 0$ and further satisfy A1-A3, the remaining heterogeneity after conditioning on Z , ζ_Z , can be approximated using traditional meta-regression. This approach targets the same parameter ϕ_{rx} as in equation (11).

In contrast, meta-regression is incapable of addressing the interventional logic—that is, what if all participants used a laptop, or if the proportion of laptop users were equal across studies. Even when the same assumptions are met (A1, A2, and B1-B3), meta-regression lacks the ability to estimate the remaining heterogeneity under hypothetical interventions. Instead, it estimates the remaining heterogeneity after additionally conditioning on target factor W' as γ_{rx} from the outcome model in equation (12).

4.2. Generalization to Nonlinear Models

In this section, we provide an imputation estimator that is flexible to accommodate nonlinear relationships or various types of variables. This type of estimator is common in mediation contexts and has been also adopted in the causal decomposition literature (e.g., Lundberg, 2020; Sudharsanan & Bijlsma, 2021; Park, Qin, & Lee, 2022). For each estimand answering Question 2 and 3, we have outlined the estimation steps using an imputation estimator.

Question 2. We estimate the remaining heterogeneity after adjusting for Z (ζ_Z) as below.

1. Fit an outcome model, regressing the outcome on R , X , and Z . Generalized linear or additive models (GLM, GAM), as well as semi- or nonparametric regression models, may be used. Linear models should be specified flexibly, incorporating polynomials and interaction terms.
2. Using the fitted outcome model from step 1, predict four potential outcomes for each individual by setting R and X to each of their possible combinations (1,1), (1,0), (0,1), and (0,0). This yields the individual level predicted outcome $\mu_{Y|1,1,Z}$, $\mu_{Y|1,0,Z}$, $\mu_{Y|0,1,Z}$, and $\mu_{Y|0,0,Z}$.
3. Then, the remaining heterogeneity is estimated as $\hat{\zeta}_Z = E[\mu_{Y|1,1,Z} - \mu_{Y|1,0,Z} - \{\mu_{Y|0,1,Z} - \mu_{Y|0,0,Z}\}]$. The heterogeneity attributable to Z is estimated as $\hat{\delta}_Z = \hat{\tau}_{rx} - \hat{\zeta}_Z$. Their standard errors are obtained as nonparametric bootstrap.

The validity of the conditional heterogeneity estimate depends on the correct specification of the outcome model given R , X , and Z .

Question 3 - static intervention. A corresponding imputation estimator is also available for the estimands in Question 3. For both estimands, it is important to mean-center all baseline covariates Z by subtracting the corresponding mean in the population from each value. This step ensures that the resulting estimates directly capture the marginal remaining heterogeneity, averaged over the population distribution of Z .

1. Fit an outcome model, regressing the outcome on R , X , Z , W , and W' . As with the estimation of the estimand in Question 2, generalized linear or additive models (GLM, GAM), as well as semi- or nonparametric regression models, may be used. In specifying the models it is important that the interactions between R , X and W' are

captured. In addition, we encourage researchers to allow Z and W to moderate the treatment effects.

2. Using the fitted outcome model from step 1, predict two potential outcomes for each individual based on two datasets that set X' and W' to $(1, w')$ and $(0, w')$. This yields the individual-level predicted outcome $\mu_{Y|R,X=1,Z,W,w'}$ and $\mu_{Y|R,X=0,Z,W,w'}$.
3. For each individual, calculate the difference between these predicted outcomes from step 2 as $\mu_{w'} = \mu_{Y|R,X=1,Z,W,w'} - \mu_{Y|R,X=0,Z,W,w'}$.
4. Finally, regress this difference $\mu_{w'}$ on R and Z . The coefficient of R from this final regression is the remaining heterogeneity $\zeta_{w'}$. The heterogeneity attributable to this static intervention is obtained as $\delta_{w'} = \hat{\zeta}_Z - \hat{\zeta}_{w'}$. Their standard error are obtained as nonparametric bootstrap.

The validity of the remaining heterogeneity estimate depends on the correct specification of the outcome model given R, X, Z, W , and W' .

Question 3 - stochastic intervention. We estimate the reduced heterogeneity and the remaining heterogeneity under a stochastic intervention of W' as below.

1. Fit a sufficiently flexible linear, generalized, or additive model, regressing the target factor W' on baseline covariates Z using only data from the reference replication study ($R = 0$). The predicted values of this model using the entire sample is denoted as $\tilde{W}' \equiv \mu_{W'|R=0,Z}$.
2. As with the static intervention, fit a sufficiently flexible outcome model, regressing the outcome Y on R, X, Z, W , and W' using the full dataset.
3. For the replication study $R = 1$, predict two potential outcomes for each individual based on two datasets that set X' and W' to $(1, \tilde{W}')$ and $(0, \tilde{W}')$. This yields the individual-level predicted outcomes $\mu_{Y|R=1,X=1,Z,W,\tilde{W}'}$ and $\mu_{Y|R=1,X=0,Z,W,\tilde{W}'}$.

4. Compute the counterfactual individual treatment effects (ITEs) for $R = 1$ group by calculating the difference between the predicted outcomes as

$$\mu_{Y|R=1, X=1, Z, W, \tilde{W}'} - \mu_{Y|R=1, X=0, Z, W, \tilde{W}'}.$$

5. Finally, regress the counterfactual ITEs on baseline covariates Z and the intercept of the regression is denoted as $\theta_{W'|1, Z}$.

6. Then, the heterogeneity attributable to this stochastic intervention is estimated as

$$\delta_{G_{W'}} = \hat{\mu}_{Y|1,1,Z} - \hat{\mu}_{Y|1,0,Z} - \hat{\theta}_{W'|1,Z} \text{ and the remaining heterogeneity is estimated as } \zeta_{G_{W'}} = \hat{\theta}_{W'|1,Z} - \{\hat{\mu}_{Y|0,1,Z} - \hat{\mu}_{Y|0,0,Z}\}.$$

Their standard errors are obtained as nonparametric bootstrap.

The imputation estimators clearly allow for a more flexible model specification than the coefficient-based estimator and, thus, are able to readily accommodate any nonlinear functional relations between variables. Nonetheless, consistent effect heterogeneity estimators are only obtained if the the models are correctly specified with regard to R, X, Z, W , and W' . Model specifications are particularly challenging because the remaining heterogeneity arises from higher-order interaction effects between the treatment, replication status, baseline covariates, and mediators, creating a significant risk of model misspecifications. While saturating the models with higher-order interactions increases flexibility, it introduces a substantial risk of overfitting, where the estimator captures noise unique to the data at hand and does not generalize well to other data. While the standard practice is to include theoretically important interactions, it is often impossible to have a well-developed theory with high-dimensional data.

To address the challenge of overfitting and potential model misspecification, we employ a machine learning based selection procedure. The Least Absolute Shrinkage Selection Operator (Lasso), introduced by Tibshirani (1996), prevents overfitting by adding a penalty to the model that constrains the coefficients. This penalty enables variable selection by shrinking the regression weight of irrelevant variables to zero. Lasso requires

tuning a hyperparameter, λ , which represents the strength of the penalty. We tune λ using cross-validation that either minimizes average validation error or chooses the simplest model within one standard error from the minimum error (James, Witten, Hastie, & Tibshirani, 2021). For both the simulation and case studies, we chose the λ that minimizes the validation error.

Despite its utility, standard Lasso is designed to minimize the prediction error but not to select confounders for causal inference (Brand, Zhou, & Xie, 2023). Specifically, it selects variables that are highly predictive of the outcome, potentially omitting variables that are needed to identify the causal effects. Failing to account for relevant covariates that are weakly correlated with the outcome but strongly correlated with the treatment leads to omitted variable bias (Belloni, Chernozhukov, & Hansen, 2014). In addition, as Lasso estimators provide a non standard sampling distribution, the standard errors cannot be reliably computed directly from the model (Tibshirani, 1996).

To mitigate these issues, we use the Double Lasso method proposed by Belloni et al. (2014). This approach involves fitting two Lasso regressions: one for the treatment and one for the outcome. We then take the union of the variables selected by both regressions and use this combined set as covariates in a final OLS regression. The final OLS model yields valid and reliable standard errors. When fitting two Lasso regressions, we include all interaction terms in addition to the main predictors. To further reduce the risk of omitted variable bias, we allow the Lasso algorithm to select from interaction terms while leaving the main predictors unpenalized. By using this approach, we robustly address the challenge of potential omitted variable bias due to regularization. Note, however, that this approach addresses this challenge when all relevant confounders of the $W' - Y$ relationship have been measured (B1).

4.3. Simulation Study

This section presents a small-scale simulation study designed to illustrate the performance of our proposed estimation methods. We compare four approaches: 1) the coefficient-based estimator, 2) the imputation estimator based on equations (11) and (12), 3) the imputation estimator combined with Double Lasso selection. The performance of these methods is assessed under conditions of both correct and incorrect model specification, across different sample sizes ($n = 500, 1000$, and 10000).

The data-generating process for our population model is as follows.

1. Generate a binary baseline covariate Z , and a treatment variable X from a Bernoulli distribution with $p = 0.5$.
2. Generate the replication status R , a mediator W , and a target variable W' based on the following models:

$$\begin{aligned} P(R = 1|Z) &= \text{logit}^{-1}(-0.5 + 0.8Z) \\ W &= 0.2 + 0.5Z + 0.7R + \epsilon_w \\ W' &= -0.3 + 0.4Z + 0.6R + 0.3W + \epsilon_{w'}, \end{aligned} \tag{13}$$

where $\epsilon_{w'}$ ϵ_w follow the standard normal distribution.

3. Finally, the outcome variable Y is generated under two different scenarios. First, the outcome is generated as

$$\begin{aligned} Y &= 0.7 + R + 0.5X + 0.3Z + 0.4W + 0.2W' + 0.1XW' + 0.5RX + 0.2RXW' \\ &\quad + 0.5XZ - 0.7RZ + 0.5RXZ + \epsilon_y, \end{aligned} \tag{14}$$

where ϵ_y follows the standard normal distribution. This outcome is used to evaluate the estimators under the assumption that the analytic model is correctly specified, that is, it coincides with the data-generating model. The second outcome-generating

model is identical but includes two additional interaction terms, $-0.5RXZ + 0.5XW$, which are omitted from the analytic model, thereby leading to model misspecification.

From the generated population data, we randomly sample with $n = 500, 1000$, and 10000 . For each sample, we estimate remaining heterogeneity using the methods described in Section 4.1 and 4.2. For the correct specified scenario, the estimation model of the coefficient-based estimator and the imputation based estimator matches the data-generating models. For the incorrect specified models, the estimation is based on equation (14), which does not include the interaction terms of X and W as well as of R and X on Z . As such, the coefficient based estimator and the imputation based estimator model are misspecified, whereas the imputation based approaches with variable selection (Double Lasso) potentially could select these interaction terms in their models. For Double Lasso, we follow the approach described in Section 4.2. We include all possible interactions of Z and W with respect to the replication study R and treatment X , allowing the Lasso algorithm to select from interaction terms. The simulation is based on 1000 iterations.

The true values for remaining heterogeneity are calculated from the population model using the definitions in Section 3. For the correctly specified models, the true value is 0.700 for a static intervention and 0.444 for a stochastic intervention on W' . For the incorrectly specified models, these true values shift to 1.050 and 0.799, respectively. This simulation study focused on estimating the remaining heterogeneity under both, static and stochastic intervention on W' . We chose these estimands because they are particularly susceptible to model misspecification as the outcome model involves the most variables (Z , W , R , X , and W').

Figure 2 summarizes the simulation results. Under correct model specification, all four methods produce asymptotically unbiased estimates. All the estimators generally perform well even with a smaller sample size ($n \leq 1000$), potentially due to a limited set of baseline covariates and mediators.

Under the incorrectly specified models, only the imputation estimator with Double

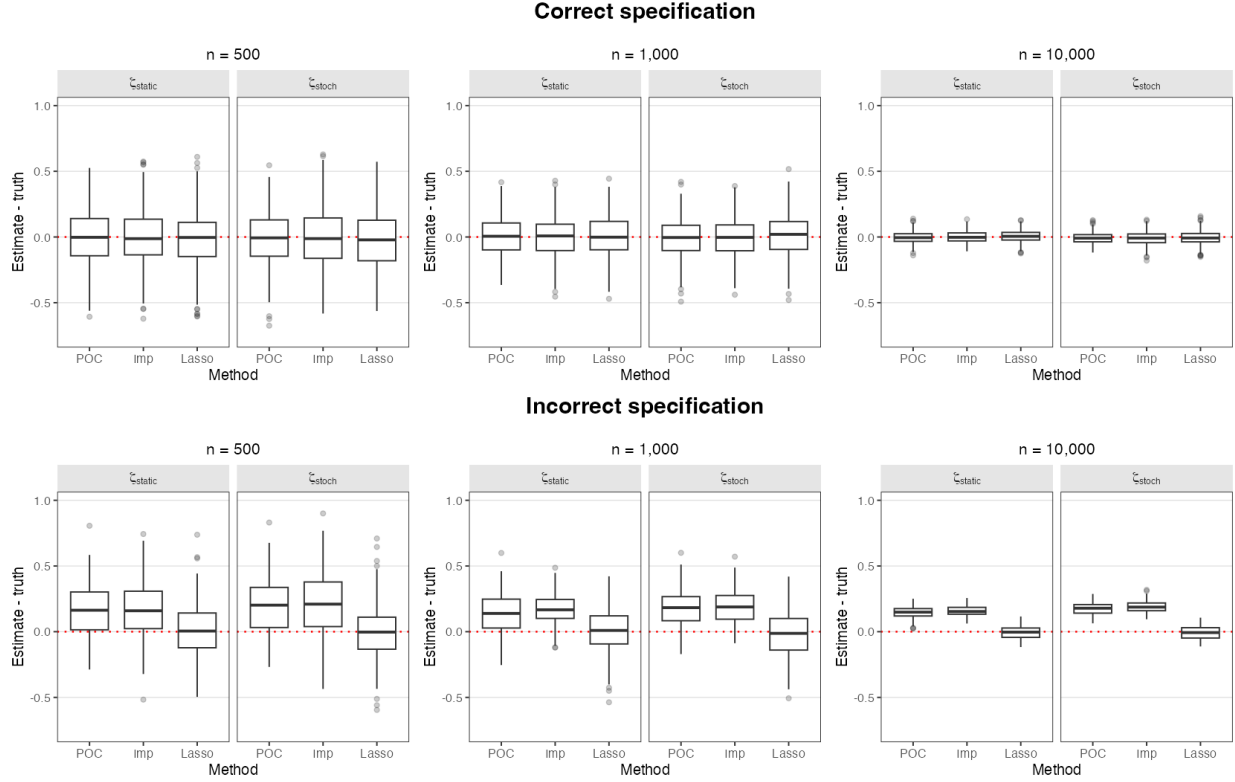


Figure 2. Simulation Results

Note: Method: 1) poc: the coefficient-based estimator, 2) imp: the imputation estimator, 3) lasso: the imputation estimator with Double Lasso selection.

Lasso method remains asymptotically unbiased. In contrast, as expected, the estimators without variable selection (methods 1 and 2) become substantially biased when the model is incorrectly specified.

Our simulation study offers practical recommendations for applied researchers. Since the coefficients-based and standard imputation estimators rely on the prespecified models defined in equations (11) and (12), they yield biased results if the data generating model deviates from these modeling assumptions. In contrast, the Double Lasso procedure selects a set of variables and interaction terms, which mitigates omitted variable bias while simultaneously preventing overfitting. Therefore, when the models are complex with numerous confounders and data sets are sufficiently large, we recommend to use the imputation estimator combined with an automated variable selection procedure, such as Double Lasso (but the Lasso needs to include sufficient higher-order terms to capture any

nonlinearity).

5. Case Study

To illustrate our approach, we applied it to replication data from a study on the imagined intergroup contact effect (Twardawski et al., 2026). The outcome was participants’ attitude toward homeless people, rated on a 1-to-101 scale ranging from ‘cold’ to ‘warm’. All four estimation approaches were applied to the data: the coefficient-based and imputation estimators based on equations (11) and (12), and the Double Lasso imputation estimator. When fitting the Double Lasso imputation estimator, we allowed baseline covariates Z and intermediate confounders W to interact with both the treatment X and replication factor R (i.e., $R * X * (Z + W)$). However, if there is a reason to believe that Z and W also interact with each other, researchers can easily allow for full interactions among all variables (i.e., $R * X * Z * W$).

Table 2 presents the respective treatment effect heterogeneity estimates. The standardized treatment effect heterogeneity (τ) is 0.258 (SE=0.107), indicating that the ATE is significantly larger in the paid provider sample than in the unpaid university sample.

As expected, the coefficient-based and imputation estimators yield almost identical results when utilizing the models defined in equations (11) and (12). However, due to the inclusion of numerous baseline covariates and their higher-order interactions, the conditional and full outcome models require the estimation of 79 and 86 parameters, respectively. With a moderate sample size of 1304, this approach potentially leads to some overfitting. The lack of parsimony is also evidenced by large standard errors compared to those produced by the Double Lasso selection procedure.

Answering Question 2, no significant effect heterogeneity remains across the three methods after adjusting for baseline covariates (participant characteristics). The remaining heterogeneity estimated using the estimators based on equations (11) and (12) and Double

Table 2
Comparison of Heterogeneity Estimates Across Three Methods

Estimand	CB	Imputation	Double Lasso
	Estimate (S.E.)	Estimate (S.E.)	Estimate (S.E.)
Question 1			
Overall Heterogeneity (τ)	0.258 (0.107)	0.258 (0.107)	0.258 (0.107)
Question 2			
Remaining Heterogeneity (ζ_Z)	0.029 (0.165)	0.029 (0.166)	0.057 (0.115)
Attributable Heterogeneity (δ_Z)	0.229 (0.145)	0.229 (0.151)	0.194 (0.119)
Question 3: static intervention			
Remaining Heterogeneity ($\zeta_{w'}$)	0.159 (0.181)	0.159 (0.170)	0.137 (0.147)
Attributable Heterogeneity ($\delta_{w'}$)	-0.130 (0.088)	-0.130 (0.101)	-0.080 (0.137)
Question 3: stochastic intervention			
Remaining Heterogeneity ($\zeta_{G_{W'}}$)	0.043 (0.164)	0.035 (0.165)	0.063 (0.128)
Attributable Heterogeneity ($\delta_{G_{W'}}$)	-0.014 (0.022)	-0.006 (0.020)	-0.006 (0.080)

Note: CB: coefficient-based estimator; Imputation: imputation estimator; Double Lasso: imputation estimator with Double Lasso using the minimum MSE penalty. Estimates statistically significant at $p < 0.05$ are indicated in bold.

Lasso method is 0.029 ($SE = 0.165$ for CB and $SE = 0.166$ for Imputation) and 0.058 ($SE = 0.107$), respectively. These results indicate that differences in participant characteristics, such as age and prior knowledge, effectively account for the observed difference in treatment effects between the two studies.

Answering Question 3, the hypothetical intervention on device type (W'), whether under a static or stochastic intervention, no significant remaining effect. While a substantive portion of the overall heterogeneity is explained by participant characteristics, the portion specifically attributable to device type is small. This suggests that the type of device used during the experiment does not meaningfully contribute to the observed effect heterogeneity after accounting for participants characteristics. This holds in both cases, whether we consider all participants conducting an experiment on a laptop (static intervention) or an equal proportion of laptop users between studies (stochastic intervention) as the comparison. As expected, the sum of the remaining and attributable

heterogeneity under both W' interventions equals the remaining heterogeneity observed after adjusting for baseline covariates.

6. Discussion

This study presents a method for identifying the drivers of treatment effect heterogeneity between replication studies. Under the potential outcomes framework, we contribute to the literature by defining, identifying, and estimating the remaining heterogeneity—both after conditioning on baseline covariates and under hypothetical interventions on a target mediator.

Our Contributions. First, we defined treatment effect heterogeneity as the observed difference in ATE between replication studies. This definition focused on the observed difference rather than the causal effect of being in a specific study. This approach is advantageous for two reasons: 1) it serves as the initial point from which to measure change in heterogeneity, and 2) it avoids the strong assumptions often required for causal inference. We then discussed which aspects of the studies (such as SUTVA) are required for a more meaningful interpretation of the treatment effect heterogeneity.

Second, we considered different approaches for disentangling drivers of treatment effect heterogeneity (Questions 2 and 3). To guide this process, we developed a DAG grounded in a real-world replication study. The DAG helped identify two potential sources of this heterogeneity: *baseline covariates* that influence replication status, and *mediators* that may vary due to differences in baseline covariates and replication status. To address Question 2, we adjusted for baseline covariates without making formal causal claims. This approach is appropriate because it is difficult to conceptualize interventions even hypothetically on socially ascribed characteristics, such as age and gender. To address Question 3, we selected target factors amenable to intervention. We utilize either a static intervention (setting the target mediator(s) to a specific value) or a stochastic intervention (equalizing the distribution of the target mediator(s) across replication studies).

Compared to Pohl et al. (in press), we narrow the scope of the causal claims, a trade-off that allows our method to rely on weaker identification assumptions. Pohl et al. (in press) investigate the causal effect of a replication factor on treatment effects by removing unintended study differences (e.g., setting variables). The approach requires measuring all relevant moderators that explain the effect heterogeneity. In contrast, our method focuses on decomposing the observed heterogeneity itself by 1) adjusting baseline covariates and 2) intervening on a well-defined single mediator. Ultimately, the choice between the methods depends on researchers' objective, whether the goal is to estimate the causal effect of a replication factor or to decompose the observed effect heterogeneity.

Third, we identified the assumptions that permit a causal interpretation of the attributable and remaining heterogeneity due to target mediators (Question 3). These assumptions are strong and are not guaranteed to hold even when the treatment is randomly assigned within each replication study. Developing a sensitivity analysis to assess the impact of violating these assumptions is beyond the scope of this study, but it is an important direction for a future study. In practice, applied researchers should make a concerted effort to enhance the plausibility of the assumptions in the context of their studies. This involves assessing baseline covariates Z and mediators W that may be confounders of the relationship of the target mediator W' and the outcome. In our example, age is a variable that strongly impacts device used and is also a possible treatment effect moderator.

Lastly, we presented and compared several estimation strategies. The coefficient-based estimator offers useful insight into the specific pathways driving the remaining heterogeneity but is limited by its reliance on restrictive pre-specified models. The imputation estimator offers more flexible alternative by accommodating nonlinear relationships and different variable types. Nonetheless, this flexibility does not eliminate the risk of model misspecification, which can introduce substantial bias. To address this challenge, we provided practical solutions that integrate variable selection into the

imputation framework, which is based on Double Lasso.

Implications for Designing Replication Studies. Several practical implications can be drawn from the proposed approach for the design of future replication studies. First, to identify different drivers of treatment heterogeneity, we recommend that researchers proactively measure both *baseline covariates* and *mediators*. The specific factors that fall into these categories depend on the replication study design. For example, when comparing two studies that differ in the timing of an experiment (summer vs. winter), baseline covariates would include participants’ prior knowledge about psychology, while a mediator would be temperature during the experiment (e.g., Twardawski et al., 2026).

Second, to identify the causal drivers of treatment heterogeneity, it is important that researchers carefully choose a target mediator. A target mediator is a factor that could potentially be manipulated and is of interest. In our example, we used device type as the target mediator, but one could also consider the location (e.g., whether participants conducted the experiment at home or in a lab (e.g., Gast et al., 2026)). The choice of the target mediator determines the specific scope of the causal conclusions. These are typically aspects, for which we aim to increase our understanding of the theory and/or which we may want to intervene on to enhance treatment effects.

Third, a causal interpretation of the heterogeneity attributable to intervening on a target mediator requires that researchers measure confounders of the relationship between the target mediator and the outcome. Note that confounders must causally precede the target mediator. For example, location (home vs. not home) may influence device choice and can therefore be a confounder for the effect of device. Conversely, device is a subsequent choice and thus cannot be a confounder for the effect of location. It is important that researchers identify and measure these confounders that can potentially moderate the treatment effect.

The proposed framework is applicable to a range of replication designs, as long as the relevant study characteristics are measured and exhibit meaningful variation across studies.

The methodology is also scalable to a case where there are more than two replication studies. In such cases, a single study is designated as the reference study, while the remaining studies are replication sites. The proposed approach can be iteratively applied to each dyadic pairing (reference vs. replication_k). By designing replication studies that carefully measure potential moderators, researchers can use our proposed framework to identify the drivers of heterogeneity in a more causally robust manner. This provides a deeper understanding of effect heterogeneity and can inform theory building.

References

- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2), 29–50.
- Brand, J. E., Zhou, X., & Xie, Y. (2023). Recent developments in causal inference and machine learning. *Annual Review of Sociology*, 49(1), 81–110.
- Cook, T. D., Steiner, P. M., & Pohl, S. (2009). How bias reduction is affected by covariate choice, unreliability, and mode of data analysis: Results from two types of within-study comparisons. *Multivariate Behavioral Research*, 44(6), 828–847.
Retrieved from <https://doi.org/10.1080/00273170903333673> doi: 10.1080/00273170903333673
- Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. Jossey-Bass.
- Doabler, C. T., Clarke, B., Kosty, D. B., Kurtz-Nelson, E., Fien, H., Smolkowski, K., & Baker, S. K. (2016). Testing the efficacy of a tier 2 mathematics intervention: A conceptual replication study. *Exceptional Children*, 83(1), 92–110.
- Feng, Y. (2024). Introduction to causal graphs for education researchers. *Asia Pacific Education Review*, 25(3), 595–609. Retrieved from <https://doi.org/10.1007/s12564-024-09980-3> doi: 10.1007/s12564-024-09980-3
- Gast, A., Höhs, J., Hoffmann, J., Kondzic, D., Pohl, S., Twardawski, M., & Sengewald, M.-A. (2026). Impact of online vs. lab setting on the evaluative conditioning effect: Identification of causal effects of replication factors. *Manuscript submitted for preparation*. doi: https://osf.io/rvms3/?view_only=4d0bc0077b0e434b84c68116e29a2cda
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). Linear model selection and regularization. In *An introduction to statistical learning: with applications in r* (pp. 225–288). Springer.
- Linden, A. H., & Hönokopp, J. (2021). Heterogeneity of research results: A new perspective from which to assess and promote progress in psychological science. *Perspectives on Psychological Science*, 16(2), 358–376. Retrieved from <https://doi.org/10.1177/1745691620964193> doi: 10.1177/1745691620964193
- Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics*, 39(1), 22–51. Retrieved from <https://doi.org/10.3102/1076998613509405> doi: 10.3102/1076998613509405
- Lockwood, J. R., & McCaffrey, D. F. (2015). Simulation-extrapolation for estimating means and causal effects with mismeasured covariates. *Observational Studies*, 1, 241–290.
- Lockwood, J. R., & McCaffrey, D. F. (2016). Matching and weighting with functions of error-prone covariates for causal inference. *Journal of the American Statistical Association*, 111(516), 1831–1839. Retrieved from <https://doi.org/10.1080/01621459.2015.1122601> doi: 10.1080/01621459.2015.1122601
- Lundberg, I. (2020). The gap-closing estimand: A causal approach to study interventions that close disparities across social categories. *Sociological Methods & Research*, 00491241211055769.
- McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large-scale replication projects in contemporary psychological research. *The American Statistician*, 73(sup1), 99–105.
- Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology*, 18(3), e3000691. Retrieved from <https://doi.org/10.1371/journal.pbio.3000691> doi: e3000691.

10.1371/journal.pbio.3000691

- Park, S. (2026). *Replication code for analyzing proxy variables and unmeasured confounding*. <https://github.com/soojinpark33/replication>. GitHub.
- Park, S., Qin, X., & Lee, C. (2022). Estimation and sensitivity analysis for causal decomposition in health disparity research. *Sociological Methods & Research*, 00491241211067516.
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (Second ed.). Cambridge: Cambridge University Press.
- Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics*, 6(2), Article 7. Retrieved from <https://doi.org/10.2202/1557-4679.1203> doi: 10.2202/1557-4679.1203
- Pohl, S., Sengewald, M.-A., Kondzic, D., Hoffmann, J., Twardawski, M., & Steiner, P. M. (in press). A causal framework for explaining effect-heterogeneity in conceptual replications. *Psychological Methods*.
- Quimpo, N. J., & Steiner, P. M. (2026). Drawing credible directed acyclic graphs for causal inference. *Psychological Methods*. Retrieved from <https://doi.org/10.1037/met0000831> (Advance online publication) doi: 10.1037/met0000831
- Schmidt, S. (2009). Shall we really do it again? the powerful concept of replication is neglected in the social sciences. *Review of general psychology*, 13(2), 90–100.
- Sengewald, M.-A., & Pohl, S. (2019). Compensation and amplification of attenuation bias in causal effect estimates. *Psychometrika*, 84(2), 589–610. Retrieved from <https://doi.org/10.1007/s11336-019-09665-6> doi: 10.1007/s11336-019-09665-6
- Sengewald, M.-A., Steiner, P. M., & Pohl, S. (2019). When does measurement error in covariates impact causal effect estimates? analytical derivations of different scenarios and an empirical illustration. *British Journal of Mathematical and Statistical*

- Psychology*, 72(2), 244–270. Retrieved from <https://doi.org/10.1111/bmsp.12146>
doi: 10.1111/bmsp.12146
- Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, 36(2), 213–236. Retrieved from <https://doi.org/10.3102/1076998610375835> doi: 10.3102/1076998610375835
- Steiner, P. M., & Kim, Y. (2016). The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. *Journal of Causal Inference*, 4(2), 1–34.
- Steiner, P. M., Shadish, W. R., & Sullivan, K. J. (2023). Frameworks for causal inference in psychological science. In H. Cooper, M. N. Coutanche, L. M. McMullen, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *Apa handbook of research methods in psychology: Foundations, planning, measures, and psychometrics* (2nd ed., Vol. 1, pp. 23–56). American Psychological Association. Retrieved from <https://doi.org/10.1037/0000318-002> doi: 10.1037/0000318-002
- Steiner, P. M., Wong, V. C., & Anglin, K. (2019). A causal replication framework for designing and assessing replication efforts. *Zeitschrift für Psychologie*, 227, 280–292. doi: <https://doi.org/10.1027/2151-2604/a000385>
- Sudharsanan, N., & Bijlsma, M. J. (2021). Educational note: causal decomposition of population health differences using monte carlo integration and the g-formula. *International Journal of Epidemiology*, 50(6), 2098–2107.
- Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., & Ellison, G. T. (2016, 12). Robust causal inference using directed acyclic graphs: the r package ‘dagitty’. *International Journal of Epidemiology*, 45(6), 1887–1894. Retrieved from <https://doi.org/10.1093/ije/dyw341> doi: 10.1093/ije/dyw341
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1), 267–288.

- Twardawski, M., Hoffmann, J., Kondzic, D., Pohl, S., Gast, A., Höhs, J., & Sengewald, M.-A. (2026). Explaining effect heterogeneity: An exemplary project on the imagined intergroup contact effect.
(Manuscript submitted for publication)
- VanderWeele, T., & Robinson, W. R. (2014). On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology (Cambridge, Mass.)*, 25(4), 473-483.
- Wong, V. C., Anglin, K., & Steiner, P. M. (2022). Design-based approaches to causal replication studies. *Prevention Science*, 23, 723–738. doi: 10.1007/s11121-021-01234-7