

The Value of Design Diversity for Knowledge Accumulation*

Anna M. Wilke[†] Cyrus Samii[‡]

March 9, 2026

Abstract

When researchers seek to accumulate evidence across settings, a challenge arises: when should study designs be harmonized, and when is design diversity valuable? We develop a decision-theoretic model in which a planner chooses between two designs of ambiguous validity using a minimax-regret criterion. We distinguish two learning goals: estimating the cross-context average effect (evidence aggregation) and detecting cross-context effect heterogeneity (external validity). Design diversity minimizes worst-case regret for aggregation because it protects against unwittingly committing to a single flawed design. Design harmonization is optimal for assessing external validity insofar as it holds design artifacts constant, isolating effect differences. These conclusions hold across alternative decision rules, analysis strategies, in a Bayesian framework, and with any number of study contexts. We demonstrate that harmonization should not be the default when designing for knowledge accumulation, and highlight trade-offs for projects that pursue both aggregation and external validity assessments.

10,348 Words

*We thank Evidence in Governance and Politics (EGAP) for support. We are grateful to Jake Bowers, Don Green, Adam Glynn, Scott Tyson, Tara Slough, Issa Dahabreh, Erin Hartman, Pablo Montagnes, James Bisbee, Michael Ting as well as to participants of at PolMeth XXXIX, WashU Formal Theory Workshop, 2023 Theory in Methods Workshop, Vanderbilt Methods Workshop, 2024 Visions in Methodology Conference, and 2025 ACIC for helpful feedback.

[†]amw703@nyu.edu, New York University

[‡]cds2083@nyu.edu, New York University

1 Introduction

The methodological turn toward causal identification has surfaced issues of validity, generalizability, and knowledge accumulation (Gerring, 2023; Imbens, 2010; Samii, 2016; Open Science Collaboration, 2015; Deaton, 2010; Deaton and Cartwright, 2018; Lucas, 2003; Thelen and Mahoney, 2015). In response, researchers make growing efforts to combine findings from several studies of the same question. Examples range from meta-analyses of organically grown research programs (see e.g. Galos and Coppock, 2023; Green and Gerber, 2019; Paluck, Green and Green, 2019*a*; Schwarz and Coppock, 2022) to coordinated initiatives that implement parallel studies in multiple populations (e.g., Banerjee, Karlan and Zinman, 2015; Blair et al., 2021; Coppock, Leeper and Mullinix, 2018; Dunning et al., 2019; Kertzer, Renshon and Xu, 2025; Slough et al., 2021). Similar efforts can be found beyond political science in economics, psychology (Michelangelo, Stephan et al., 2014; Milkman et al., 2021), and biomedical science (Park et al., 2019).

However, the literature on research design for cumulative learning remains sparse. A difference across knowledge accumulation efforts is the degree of research design harmonization. Organically grown research programs feature studies drawing on a variety of measurement protocols and treatment variations. In contrast, coordinated initiatives often design a set of studies, typically all experiments, to resemble each other as much as possible.

We use a decision-theoretic framework to shed light on when research design harmonization across studies is preferable to design diversity and vice versa. Our results speak to both design of prospective studies and to criteria for including studies in meta-analyses of existing work. An influential set of papers highlights the benefits of design harmonization (Slough and Tyson, 2023, 2024), an insight that has shaped applied knowledge accumulation efforts (e.g., Bassan-Nygate et al., 2025; Blair et al., 2021). Our contribution is to clarify the broad inferential value of research design diversity and to show that harmonization is helpful only under a specific set of conditions.

To start, we model a decision-maker who runs two studies, each in a different context, to

learn about the effect of some treatment on some outcome. This effect could be heterogeneous across contexts. For concreteness, it is helpful to think about these studies as experiments, but our analysis applies to observational studies as well. We distinguish between two common learning objectives. First, the decision-maker may want to estimate the cross-context average effect, for example to approximate some population-level mean effect as is typical in meta-analyses. We refer to this objective as *evidence aggregation*. Second, the decision-maker may want to assess the cross-context *external validity* of effects by estimating the cross-context difference in effects. Researchers often pursue both objectives, but distinguishing between them reveals trade-offs.

For each study, the decision-maker chooses between two research designs, by which we mean operationalizations of treatments and outcomes. These operationalizations may not perfectly capture the conceptual treatments and outcomes of interest and therefore cause estimates to systematically differ from the effects of interest. For example, a survey measure of sensitive behavior may suffer from experimenter demand effects. Our decision-maker faces *ambiguity* over these design artifacts. Ambiguity, sometimes referred to as Knightian uncertainty, is the “lack of knowledge of an objective probability distribution,” a condition that is common in the social sciences (Manski, 2000). The decision-maker does not know which design introduces the greater artifact, but deems a range of artifacts plausible for both designs. We intentionally suppress a priori differences in design quality to home in on the relative merits of design harmonization and diversity as distinct from quality concerns.

The question of interest is under what conditions the decision-maker will find it optimal to harmonize, i.e., prefer the same design in each study, and when she will want to diversify i.e, favor a different design in each study. Importantly, a decision-maker who uses the same design across contexts may still adjust treatment and outcome operationalizations to local conditions. The key to our notion of design harmonization is that using the same design across contexts ensures that design artifacts will be similar as well. Conversely, the design diversity we consider does not stem from ways in which researchers might tailor the same

version of a treatment or outcome measure to different contexts, but from a deliberate decision to rely on different operationalizations with distinct consequences for the validity of one’s estimates across studies.

To capture concerns about “robustness,” our decision-maker relies on a minimax regret criterion, which can define choices under ambiguity. The minimax regret approach is a robust alternative to Bayesian optimization, which requires commitment to specific prior and outcome distributions that, for social science applications, are typically disputable. Minimax regret analyses are becoming increasingly common in statistical decision theory for social science (Manski, 2004; Zhang, Huang and Imai, 2024; Dominitz and Manski, 2017; Olea, Qiu and Stoye, 2023). Regret is defined as the difference between the utility that a decision-maker could obtain if she knew the quantity of interest – here the cross-context average or difference in treatment effects – and the utility that she expects to earn given that she must rely on an estimate of this quantity. The idea is that a decision-maker chooses procedures – here whether to harmonize designs – to minimize the regret that obtains when the quantities that the decision-maker cannot control – here treatment effects and design artifacts – take the values least favorable to the decision-maker. A minimax regret criterion is typically preferred to a maximin welfare one that tends to be overly conservative (Manski, 2004).

The decision-maker’s learning objective drives whether research design harmonization or diversity is optimal. If the decision-maker seeks to estimate the cross-context average effect, using diverse designs guards against the worst case scenario in which a large design artifact compromises the validity of all estimates. However, if the decision-maker seeks to estimate the cross-context difference in effects, design harmonization holds design artifacts constant across studies, isolating treatment effect heterogeneity from design artifact differences. Researchers who are interested in both aggregating evidence *and* assessing external validity face a trade-off. Design diversity is optimal for one objective but harmonization for the other.

We then show that these results are robust to several extensions. First, our results remain identical if we allow the decision-maker to select not only whether to harmonize but also the

decision-rule to apply to the resulting estimates. Second, diversification remains the preferred approach if a decision-maker aggregates evidence not by estimating the cross-context average effect but by using a joint test of null hypotheses in each context. Third, one may worry that the minimax regret criterion, due to its focus on the worst case, stacks the deck in favor of diversification. We derive the same results in a fully Bayesian analysis which requires the addition of priors but allows for an expected utility maximizing rather than maximal regret minimizing decision-maker. Fourth, we show that our conclusions hold when artifacts of the same design are only imperfectly correlated across contexts. Fifth, the same logic applies with more than two contexts, although the availability of multiple contexts enables the decision-maker to induce some diversity without jeopardizing her ability to make certain external validity assessments.

We also develop extensions that go beyond a focus on validity or cross-context learning. First, we consider a decision-maker who faces ambiguity about the variability rather than the validity of research designs and show that research design diversity becomes the preferred choice even for external validity assessments. Second, we explore the conditions under which within-study design diversity can be helpful for the estimation of average treatment effects.

Our paper is closely related to Slough and Tyson (2023) and Slough and Tyson (2024). The core difference is that we define treatment effects independently from research designs, which introduces considerations of design validity. Slough and Tyson abstract from design-related validity problems within a context and treat the difference in measured outcomes induced by the chosen treatment-control contrast as the inferential target. Since Slough and Tyson’s framework does not allow for multiple research designs that target the same estimand, harmonization has only upsides. Harmonization can be beneficial in our framework as well, but only under certain conditions. While harmonization can be advantageous for external validity assessments, design diversity dominates for the purposes of evidence aggregation because it bestows robustness.¹

¹Slough and Tyson (2024) informally discuss the advantages of design variations albeit for a different reason, namely to learn about design artifacts.

Our main lesson for empirical practice is that researchers who seek to accumulate knowledge should not default to harmonization, but instead justify their design choices in terms of their inferential concerns and objectives. Design diversity among existing studies is not necessarily an obstacle for meta-analyses that seek to estimate average effects. Our results also highlight the benefit of partial harmonization – including both harmonized and diverse design elements across studies – as a way to navigate the trade-off between knowledge accumulation and external validity assessments. We illustrate these insights by discussing two recent knowledge accumulation efforts (Bassan-Nygate et al., 2025; Blair et al., 2021).

A robustness rationale for research design diversity has been discussed in various literatures including in informal work on “conceptual” replications (Kaelin Jr, 2017; Nosek and Errington, 2017, 2020). Empirical evidence from genetics research finds that results replicate better if produced by decentralized research communities (Danchev, Rzhetsky and Evans, 2019). “Stimulus sampling” theory proposes that researchers vary treatment operationalizations to increase confidence in both validity and generalizability (Wells and Windschitl, 1999; Monin and Oppenheimer, 2014). Callis, Dunning and Tuñón (2023) suggest a diversification strategy to learn about the mechanisms through which treatments operate. Research on experimental design advocates for the use of diverse placebo conditions within a single experiment (Porter and Velez, 2022). Work on external validity (Egami and Hartman, 2020) and measurement (Fu and Green, 2025) points to robustness gains from the inclusion of multiple outcome measures within the same study. Rosenbaum (2010) shows that multiple plausibly biased observational designs can yield more robust conclusions than any on its own. We formalize such intuitions for the case of cross-context research design.

This paper adds to a growing literature that uses formal tools to study optimal research design (Abramson, Koçak and Magazinnik, 2022; Azevedo et al., 2020; Banerjee, Chassang and Snowberg, 2017; Banerjee et al., 2020; Kasy, 2016; Izzo, Dewan and Wolton, 2018; Little and Pepinsky, 2021; Recht, 2025). Setups similar to ours have also been used to study optimal publication rules (Frankel and Kasy, 2022), pre-registration requirements (Kasy and Spiess,

2022) and the value of significance testing (Abadie, 2020).

2 Illustrating key concepts

We are interested in how to design or select studies of a treatment across multiple contexts to gain cumulative knowledge (e.g., Dunning et al., 2019; Blair and McClendon, 2021). Our framework speaks to designing primary studies de novo or a meta-analysis of existing studies. We illustrate our framework using the interdisciplinary research program on “inter-group contact” as an example. This literature studies the effects of interpersonal contact between groups on inter-group relations and has spurred several meta-analyses.² The idea, going back to Allport (1954), is that interactions with out-group members can reduce negative dispositions towards out-groups.

2.1 Learning objectives

We presume that cross-context research efforts aim to shed light on a causal mechanism like the one depicted by the directed-acyclic graph (DAG) in panel (i) of Figure 1.³ A mechanism consists of a treatment T that may affect an outcome Y , where this relationship may be modified by individual-level or contextual background characteristics X . T could represent contact with out-group members, Y behaviors towards the out-group and X the pre-existing level of inter-group conflict.

We define two learning objectives that researchers may have regarding such a mechanism. First, researchers may aim to summarize the distribution of effects of, say, inter-group contact across studied contexts using a statistic like the average. We call this goal *evidence aggregation*. In meta-analyses, researchers typically assume the average effect across studied contexts corresponds to the average effect across some broader population (e.g., Gelman et al., 2013,

²See Lowe (2025) Pettigrew and Tropp (2006), Pettigrew et al. (2011), Paolini, Harwood and Rubin (2021), Paluck and Green (2009), Paluck, Green and Green (2019b), and Vezzali et al. (2018).

³See Humphreys and Jacobs (2023, pp. 30-67) for more on how to depict mechanisms using DAGs.

chap. 5.6). The literature on transportability shows how (weighted) averages of treatment effects across studied contexts can be used to approximate effects in unstudied contexts (e.g., Egami and Lee, 2024). A social-welfare-maximizing decision-maker who weights contexts equally may base rollout decisions on the average treatment effect across contexts.

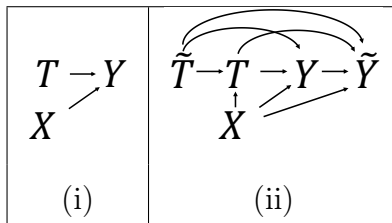


Figure 1: Directed acyclic graphs characterizing (i) a mechanism connecting a conceptual treatment T to a conceptual outcome Y , with the effect potentially moderated by a background factor X , and (ii) ways we may observe this mechanism using operationalizations of the treatment \tilde{T} and outcome \tilde{Y} .

A second goal is to assess cross-context *external validity*.⁴ Researchers may want to understand whether conclusions about the effects of inter-group contact generalize from peaceful to conflict-active settings. Here, the inferential target is the cross-context difference in effects. This goal has been a major motivation for multi-site studies (Blair and McClendon, 2021). Slough and Tyson (2024) provide ways to test whether treatment effects differ across contexts. As a welfare criterion, the cross-context difference matters if a decision-maker can implement an intervention in only a subset of contexts and seeks to choose the context where the intervention produces the greatest welfare gain.

2.2 Research design

While broader notions of research design exist (Blair et al., 2019), we follow existing work on harmonization (Slough and Tyson, 2023) and conceptualize a research design as a treatment

⁴We focus on variation in effects across *contexts*, as this is the external validity dimension of substantive interest in many knowledge accumulation settings. In principal, our framework could be extended to study external validity with regard to variations in treatments, outcomes, or contexts (Campbell, Stanley and Gage, 1966; Egami and Hartman, 2023).

operationalization and an outcome measurement strategy. Our primary way of thinking about design thus concerns operationalizations of variables, not causal identification strategies.

Researchers use these operationalizations as a means to draw inferences about a mechanism like the one depicted in Figure 1. One way to think about the distinction between the target of inference in a given context and a research design is as the difference between the effect of an idealized treatment on some real-world outcome and the effect of some practically implementable treatment on an imperfect measure of this outcome. The idealized treatment could be a friendly conversation with an out-group member each week that lasts for a minimum of one hour. A practical intervention may be assignment to mixed-group computer classes. The real-world outcome may be whether someone discriminates against out-group members. Possible outcome measures may be survey questions or dictator games.

A second way to think about the target-design distinction is as the difference between the quantity that is decision-relevant for a policy-maker and the quantity that can be studied. Rarely is it feasible to study the effects of all interventions in the exact shape in which policy-makers would like to implement them on the actual behaviors that policy-makers care about. Hence, policy-makers have to think of the particular treatment operationalization as tied to the conceptual class of interventions that may be rolled out at scale and of the outcome measures as tied to the real-world outcomes of interest.

That the same conceptual outcome or treatment can be operationalized in multiple ways is uncontroversial. Table 1 shows how researchers have operationalized inter-group contact in recent field experiments, including the formation of heterogeneous sports teams and educational programs. Heterogeneous teams have played different sports and educational programs have taught varying skills. Individuals interacted for different periods of time and with varying shares of out-group members. Similar diversity exists for outcome measures, as shown in Table 2. Behavior towards out-group members has been captured through various behavioral measures and survey items. Some behavioral measures are collected in lab-like settings while others record real-world behavior.

	Category	Activity	Duration	% out-group per team/classroom
(Mousa, 2020)	Sports	Soccer	2 months	33%
(Lowe, 2025)	Sports	Cricket	<1 month	Randomly drawn share of 5
(Scacco and Warren, 2018)	Education	Computer training	4 months	40% - 60%
(Zhou and Lyall, 2025)	Education	Various TVET courses	3 and 6 months	34%-66%

Table 1: Example operationalizations of inter-group contact as a treatment

	Type of Measure	Details
(Mousa, 2020)	Behavioral	Redeeming a voucher for a restaurant in an out-group neighborhood
	Behavioral	Vote for out-group soccer player to receive sportsmanship prize
	Behavioral	Attend mixed dinner event
	Behavioral	Donate \$1 survey compensation to neutral versus in-group NGO
	Survey question	Self-reported frequency of training soccer with out-group members
	Survey question	Willingness to register for mixed soccer team in the future
	Survey question	Self-reported comfort with having Muslim neighbors
(Scacco and Warren, 2018)	Behavioral	Dictator game with out- and in-group members
(Zhou and Lyall, 2025)	Survey question	Self-reported level of interaction with migrants outside of treatment

Table 2: Example operationalizations of behavior towards out-groups

In practice, it can be difficult to determine whether distinct outcome measures and treatment variations capture the same concept. We do not attempt to solve this problem here but believe the answer will depend on theoretical considerations.

Panel (ii) of Figure 1 represents our notion of research design by embedding the mechanism in a more complicated DAG that involves a treatment operationalization \tilde{T} and outcome operationalization \tilde{Y} . The arrows linking \tilde{T} to Y and \tilde{Y} and linking T to \tilde{Y} are causal relationships induced by the research design. These design artifacts lie outside the causal mechanism of interest. They can stem from classic internal validity concerns and from incidental ways in which the chosen outcome and treatment operationalizations do not resemble the conceptual outcome and treatment of interest – what Cronbach and Meehl (1955) call low construct validity.

For example, dictator game behavior may proxy for workplace discrimination against

out-group members (as indicated by the arrow from Y to \tilde{Y}), but behavior in a low-stakes lab setting may be more or less sensitive to inter-group contact than the real-world behaviors of interest. Survey items can ask directly about these behaviors but may suffer from social desirability bias. Contact with out-group members may cause individuals to see positive behavior towards out-groups as socially desirable and to report engaging in such behavior even if they would not in fact do so. There may thus be an effect of inter-group contact (T) on survey response (\tilde{Y}) in the absence of an effect on actual behavior (Y).

Treatment operationalizations may induce unrealistically high levels of contact that could not be achieved outside an experiment or may fail to create contact (as indicated by the arrow from \tilde{T} to T). A mixed classroom intervention, say, may struggle to induce contact if students discriminate against out-group members when choosing study partners. Moreover, many contact treatments expose participants to additional experiences beyond contact (e.g., educational content). These ancillary components may have direct effects on behavior towards out-groups that do not go through inter-group contact (the arrow from \tilde{T} to Y) – an exclusion restriction violation (Morton and Williams, 2010, chapt. 7). Some ways to introduce inter-group contact may create experimenter demand effects. Lowe (2025) finds online contact produces greater effect estimates than other contact interventions and points to demand effects as a potential explanation.

These examples illustrate how treatment and outcome operationalizations may compromise validity. Research designs in our framework differ in terms of the artifacts they produce, and researchers are uncertain about the quality of each design. A researcher may not know which measure is worse and how much each measure will cause estimates to deviate from the quantity of interest. In this scenario one would ideally use both measures in any given study. To home in on the trade-offs inherent in design harmonization and diversity, we assume a single design must be chosen for each study. Implicitly, we imagine a decision-maker facing a budget constraint. This idea seems intuitive for treatment variations, which are typically costly to implement. Even for outcome measures, constraints exist in terms of cost,

questionnaire space and respondents' attention span.

We focus on a situation in which a researcher chooses between two designs that are *ambiguity-equivalent*, i.e., for which she deems the same range of artifacts plausible. We suppress quality differences across designs, not because we think this assumption always applies but to home in on the up- and downsides of research design harmonization and diversity as distinct from quality considerations.

2.3 Harmonization

Our core notion of harmonization is that using the same design across studies introduces dependence in design artifacts. For clarity of exposition, we begin by assuming that this dependence is perfect, i.e., if two studies are harmonized with the same design, they have the same artifact. Later, we analyze the more general case in which harmonization introduces imperfect cross-context dependence in artifacts. Our results remain identical as long as artifacts of the same design are not completely independent. Once a given design produces completely independent artifacts across contexts, harmonization becomes meaningless.

We presume that even researchers who harmonize adjust operationalizations to each context, e.g., by translating surveys or adjusting names and examples. Allowing for such adaptation is crucial to make the assumption that design artifacts remain similar across contexts plausible. Imagine an extreme scenario in which a researcher fails to translate a survey question used in country A to the language spoken in country B . Design artifacts are going to vary across countries A and B if respondents in country A do and those in country B do not understand the question. Harmonization here means using the same operationalizations across contexts, but up to changes required to adapt the operationalizations to local conditions.

3 Formal setup

There are two study contexts $i \in \{1, 2\}$, though we generalize to more contexts below. A decision-maker faces a binary choice $a \in \{0, 1\}$ regarding some treatment, e.g., whether or where to implement the treatment, or whether to report in a journal article that the treatment has a positive effect. We provide more detail on how to interpret a below. $\tau_i \in \mathbb{R}$ is the average effect of the treatment in context i . τ_i may be the effect of some idealized treatment of scientific interest on real-world behavior or simply the decision-relevant quantity for some policy-maker. Our analysis does not require us to model the units in each context i , but the reader may imagine that each context i contains n_i units. τ_i then is the average across the unit-level treatment effects for all n_i units in context i . We return to the decision-maker’s objective below as well. For now, it suffices to say that the decision-maker’s optimal decision a^* depends on the vector $\boldsymbol{\tau} = (\tau_1, \tau_2)$ of average treatment effects. The decision-maker does not observe $\boldsymbol{\tau}$ but is motivated to learn about it to improve her choice a . The decision-maker conducts two studies, one in each context.

For each study, the decision-maker chooses one of two possible research designs $j \in \{1, 2\}$, where we think of research designs as treatment and outcome operationalizations. A study in context i that makes use of design j produces an estimate of $\hat{\tau}_i^j$ of the effect τ_i . For example, $\hat{\tau}_i^j$ may be a difference in means obtained from an experiment run on a sample of the n_i units in context i . $\hat{\tau}_i^j$ may differ from τ_i for systematic and for statistical reasons. Specifically,

$$\hat{\tau}_i^j - \tau_i = \delta_i^j + \epsilon_i.$$

δ_i^j represents a “design artifact” introduced by design j in context i and ϵ_i represents mean-zero statistical noise from random assignment or sampling such that $\mathbb{E}[\hat{\tau}_i^j] = \tau_i + \delta_i^j$. As such, δ_i^j captures the degree of bias relative to τ_i and therefore the validity of research design j (Nosek et al., 2022).⁵ If $\delta_i^j = 0$, design j produces valid estimates of τ_i , i.e., $\mathbb{E}[\hat{\tau}_i^j] = \tau_i$. If $\delta_i^j \neq 0$,

⁵Imai, King and Stuart (2008) offer a linear error decomposition that is similar in spirit but focuses

then the estimate distribution produced by design j is centered on a quantity that differs from the effect τ_i of interest. Design artifacts may arise if the chosen operationalizations do not adequately capture the outcomes and treatments of interest. For example, research design j may consist of an outcome measure that suffers from treatment-related measurement error. If experimenter demand leads subjects in the treatment group to over-report socially desirable behavior, effect estimates will be too large, i.e., $\delta_i^j > 0$. Appendix C provides one possible formalization of design artifacts using unit-level potential outcomes.

The decision-maker faces ambiguity over the quality of the research designs available to her, meaning she does not know the objective probability distribution governing the artifacts. We presume that $\delta_i^j \in [\underline{\delta}_i, \bar{\delta}_i]$ for $j \in \{1, 2\}$ with $\underline{\delta}_i, \bar{\delta}_i \in \mathbb{R}$ and $\underline{\delta}_i < \bar{\delta}_i$ for $i \in \{1, 2\}$. Each design introduces an artifact from within a plausible range that is identical across both designs within a given context i . In that sense, the two designs are *ambiguity-equivalent*. We also assume that $|\delta_i^j - \delta_{i'}^{j'}| \geq \Delta > 0$ for all i and i' . We rule out that the two designs introduce identical artifacts to avoid a knife-edge case in which the decision-maker is indifferent between design harmonization and diversity. The event that the two artifacts are identical would occur with zero probability if they were random variables.

In our first analysis, we make two simplifying assumptions. These assumptions bring the logic of our analysis into sharp relief. Later, we show that loosening these assumptions does not change the fundamental logic, but requires additional notation that somewhat obscures what the simplified analysis clarifies. The first simplifying assumption is to fix $\delta_i^j = \delta^j$, $\underline{\delta}_i = \underline{\delta}$, and $\bar{\delta}_i = \bar{\delta}$ across contexts. Put differently, we first assume a given design j introduces the same artifact across contexts and that the largest and smallest possible value that this artifact can take does not vary with context as well. Later, we show that the same conclusions can be derived from a more flexible model that varies the degree of dependence in the artifacts introduced by a given design across contexts.

on biases due to sample selection and confounded treatment assignment, bracketing any consideration of treatment or measurement validity. By contrast, our decomposition brackets consideration of sample selection and confounding, and focuses on validity of treatment or outcome measures for targeting the decision-relevant effect.

The second simplifying assumption is that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for all i . The normality assumption can be justified with reference to the finite population central limit theorem (CLT) (Li and Ding, 2017). The assumption that the variance σ^2 is constant across designs reflects that the two designs are equally powered. Below, we show that our main conclusion is strengthened if different designs may produce estimates of varying precision and the decision-maker faces ambiguity over the variability of each design.

The question of interest is whether the decision-maker, given different learning objectives, finds it optimal to use the same or different research designs across the two studies. We denote the choice to harmonize, i.e., to use the same design, by $h \in \{0, 1\}$. Let $\boldsymbol{\delta}$ be a two-dimensional vector with the i th element equal to the design artifact of the study in context i . Since research designs differ only in terms of their artifacts, the choice between design diversity and harmonization is a choice of $\boldsymbol{\delta}$. We assume w.l.o.g. that harmonization means using design 1 for both studies, while diversity means using design 1 in context 1 and design 2 in context 2:

$$\boldsymbol{\delta}(h) = \begin{cases} (\delta^1, \delta^1) & \text{if } h = 1, \\ (\delta^1, \delta^2) & \text{if } h = 0. \end{cases} \quad (1)$$

The decision-maker first chooses whether to harmonize and then runs both studies. Next, she observes the vector of estimates $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1^j, \hat{\tau}_2^j)$ but not the vector of average treatment effects $\boldsymbol{\tau}$, the vector of design artifacts $\boldsymbol{\delta}(h)$, nor the vector of error terms $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2)$. Finally, the decision-maker chooses a .

If the decision-maker chooses $a = 0$, she receives a status quo payoff that we normalize to zero. If she chooses $a = 1$, her utility is given by a linear functional $f(\cdot)$ of the average

treatment effects $\boldsymbol{\tau}$:

$$u(a; \boldsymbol{\tau}) = \begin{cases} 0 & \text{if } a = 0, \\ f(\boldsymbol{\tau}) & \text{if } a = 1. \end{cases} \quad (2)$$

This objective function implies that the decision-maker would like to choose $a = 1$ if and only if $f(\boldsymbol{\tau})$ exceeds zero. Hence, she is motivated to learn about $f(\boldsymbol{\tau})$. This formulation allows us to capture the two learning objectives previewed above. First, we consider $f(\boldsymbol{\tau}) = \frac{\tau_1 + \tau_2}{2}$, i.e., the decision-maker’s estimand is the cross-context average in treatment effects. We refer to this case as *evidence aggregation*. Second, we consider $f(\boldsymbol{\tau}) = \tau_1 - \tau_2$, i.e., the decision-maker’s estimand is the cross-context difference in treatment effects.⁶ We refer to this case as learning about *external validity*.

The need to choose a creates the decision-maker’s learning incentive. Substantively, a may represent a researcher’s choice of whether to report that the estimand $f(\boldsymbol{\tau})$ exceeds zero in a journal article, assuming researchers’ gains from doing so are proportional to the true quantity of interest. Alternatively, we may think about a welfare-motivated policy-maker who chooses whether to implement the treatment under study in both contexts (for the cross-context average) or in context 1 rather than context 2 (for the cross-context difference), where welfare in context i equals τ_i . We derive identical results from a Bayesian analysis with a squared error loss function.

Given her ambiguity about treatment effects and design artifacts, the decision-maker uses a minimax regret criterion. Let $a^* = \arg \max_{a \in \{0,1\}} u(a; \boldsymbol{\tau})$ be the decision-maker’s optimal choice and $u(a^*; \boldsymbol{\tau})$ the resulting maximal utility given the vector of average treatment effects

⁶Identical results obtain if $f(\boldsymbol{\tau}) = \tau_2 - \tau_1$.

$\boldsymbol{\tau}$. It follows from equation 2 that

$$a^* = \begin{cases} 1 & \text{if } f(\boldsymbol{\tau}) > 0, \\ 0 & \text{if } f(\boldsymbol{\tau}) \leq 0. \end{cases} \quad (3)$$

Because the decision-maker does not observe $\boldsymbol{\tau}$, she is unable to directly implement a^* and cannot guarantee herself the payoff $u(a^*; \boldsymbol{\tau})$. Instead, the decision-maker needs to base her decision a on the vector of estimates $\hat{\boldsymbol{\tau}}$. We assume that the decision-maker computes $f(\hat{\boldsymbol{\tau}})$ and chooses a according to the following decision rule:

$$a = \begin{cases} 1 & \text{if } f(\hat{\boldsymbol{\tau}}) > 0, \\ 0 & \text{if } f(\hat{\boldsymbol{\tau}}) \leq 0. \end{cases} \quad (4)$$

This assumption is natural for several reasons. First, it is common in the literature on minimax regret to restrict attention to empirical success rules which “emulate[s] the optimal treatment rule by replacing unknown response distributions with sample analogs” (Manski, 2004, p. 1231). Second, since $f(\cdot)$ is linear, using $f(\hat{\boldsymbol{\tau}})$ as an estimator of $f(\boldsymbol{\tau})$ does not produce any additional bias beyond that which stems from the existence of design artifacts. In an extension, we show that our results are robust to a model in which the decision-maker optimally chooses a cutoff c on $f(\hat{\boldsymbol{\tau}})$ above which she implements the treatment.

The decision-maker’s regret $r(\boldsymbol{\tau}, \boldsymbol{\delta}(h))$ is defined as the difference between the decision-maker’s maximal utility $u(a^*; \boldsymbol{\tau})$ and the utility she expects to earn given that she makes her choice a according to equation 4:

$$r(\boldsymbol{\tau}, \boldsymbol{\delta}(h)) = u(a^*; \boldsymbol{\tau}) - \mathbb{E}_{\boldsymbol{\epsilon}}[u(a; \boldsymbol{\tau})]. \quad (5)$$

$f(\hat{\boldsymbol{\tau}})$ may differ from $f(\boldsymbol{\tau})$ due to design artifacts $\boldsymbol{\delta}(h)$ and statistical noise $\boldsymbol{\epsilon}$. The regret function takes the vector of average treatment effects $\boldsymbol{\tau}$ and design artifacts $\boldsymbol{\delta}(h)$ as given.

The expectation in the second part of the expression is taken only over the vector of error terms ϵ . The decision-maker’s regret is a function of her harmonization choice h , since this choice determines each study’s design artifact and hence shifts the location of the distribution of estimates $f(\hat{\tau})$.

The decision-maker considers the maximum regret over all possible vectors τ of treatment effects and vectors $\delta(h)$ of design artifacts, where the latter depends on her harmonization choice. She chooses whether to harmonize to minimize this maximum regret:

$$\min_h R(h), \text{ where } R(h) := \max_{\tau, \delta(h)} r(\tau, \delta(h)). \quad (6)$$

This criterion allows her to make a choice that is robust in that it improves the most regrettable outcome over the range of possible artifacts and treatment effects.

4 Analysis

The decision-maker wants to choose $a = 1$ if and only if the estimand is positive, that is, if $f(\tau) > 0$. For example, a researcher wishes to report that the average effect of inter-group contact is positive, or that inter-group contact is more effective in one context than in another, only when these statements are true. If the decision-maker observed $f(\tau)$ directly, she would always choose optimally. In practice, however, she must rely on her estimate $f(\hat{\tau})$, which is subject to sampling uncertainty and design artifacts.

Regret arises in two cases. If the estimand is positive, regret occurs when the estimate fails to exceed zero and the decision-maker incorrectly chooses $a = 0$. If the estimand is negative, regret occurs when the estimate exceeds zero and the decision-maker incorrectly chooses $a = 1$. Across both cases, regret is driven by the probability that the estimate crosses the decision threshold – here, zero – in the wrong direction. Figure 2 below illustrates this logic for the case of a positive estimand.

Design artifacts shift the location of the estimate distribution away from the quantity of

interest. The magnitude of this shift is $f(\boldsymbol{\delta}(h))$, which depends both on the decision-maker's learning objective and on her harmonization choice h . In the evidence aggregation case, the shift equals the cross-study average of design artifacts; in the external validity case, it equals the cross-country difference in artifacts. The artifact vector $\boldsymbol{\delta}(h)$ itself is determined by the harmonization choice h .

Whether such shifts increase or decrease regret depends on the true state of the world. If the estimand is positive, upward shifts in the estimate distribution reduce regret by increasing the likelihood that the decision-maker correctly chooses $a = 1$. If the estimand is negative, the same upward shifts increase regret by making the incorrect choice of $a = 1$ more likely. The opposite holds for downward shifts.

The decision-maker considers the maximum regret that could arise across all possible values of the estimand and all admissible realizations of design artifacts implied by her harmonization choice. She decides whether to harmonize or diversify to minimize this maximum regret.

Let $\boldsymbol{\delta}^{\min}(h)$ and $\boldsymbol{\delta}^{\max}(h)$ denote the realizations of the artifact vector that respectively minimize and maximize $f(\boldsymbol{\delta}(h))$. Thus, $f(\boldsymbol{\delta}^{\min}(h))$ and $f(\boldsymbol{\delta}^{\max}(h))$ represent the smallest and largest possible shifts in the location of the estimate distribution that can arise under harmonization choice h . Lemma 1 characterizes the decision-maker's worst-case regret in terms of these shifts; all proofs are provided in the appendix.

Lemma 1. *The decision-maker's choice problem can be written as*

$$\min_h R(h), \text{ where } R(h) = \max_{\boldsymbol{\tau}} \tilde{r}(\boldsymbol{\tau}, h) = \begin{cases} f(\boldsymbol{\tau})\Phi\left(\frac{-f(\boldsymbol{\tau})-f(\boldsymbol{\delta}^{\min}(h))}{\mathcal{B}_{f(\cdot)}\sigma}\right) & \text{if } f(\boldsymbol{\tau}) > 0 \\ -f(\boldsymbol{\tau})\Phi\left(\frac{f(\boldsymbol{\tau})+f(\boldsymbol{\delta}^{\max}(h))}{\mathcal{B}_{f(\cdot)}\sigma}\right) & \text{if } f(\boldsymbol{\tau}) \leq 0, \end{cases} \quad (7)$$

where $\mathcal{B}_{f(\cdot)=(x_1+x_2)/2} = \frac{1}{\sqrt{2}}$ and $\mathcal{B}_{f(\cdot)=x_1-x_2} = \sqrt{2}$. There exists $\boldsymbol{\tau}^* \in \mathbb{R}^2$ such that $\tilde{r}(\boldsymbol{\tau}^*, h) = \max_{\boldsymbol{\tau}} \tilde{r}(\boldsymbol{\tau}, h)$, but $\boldsymbol{\tau}^*$ need not be unique.

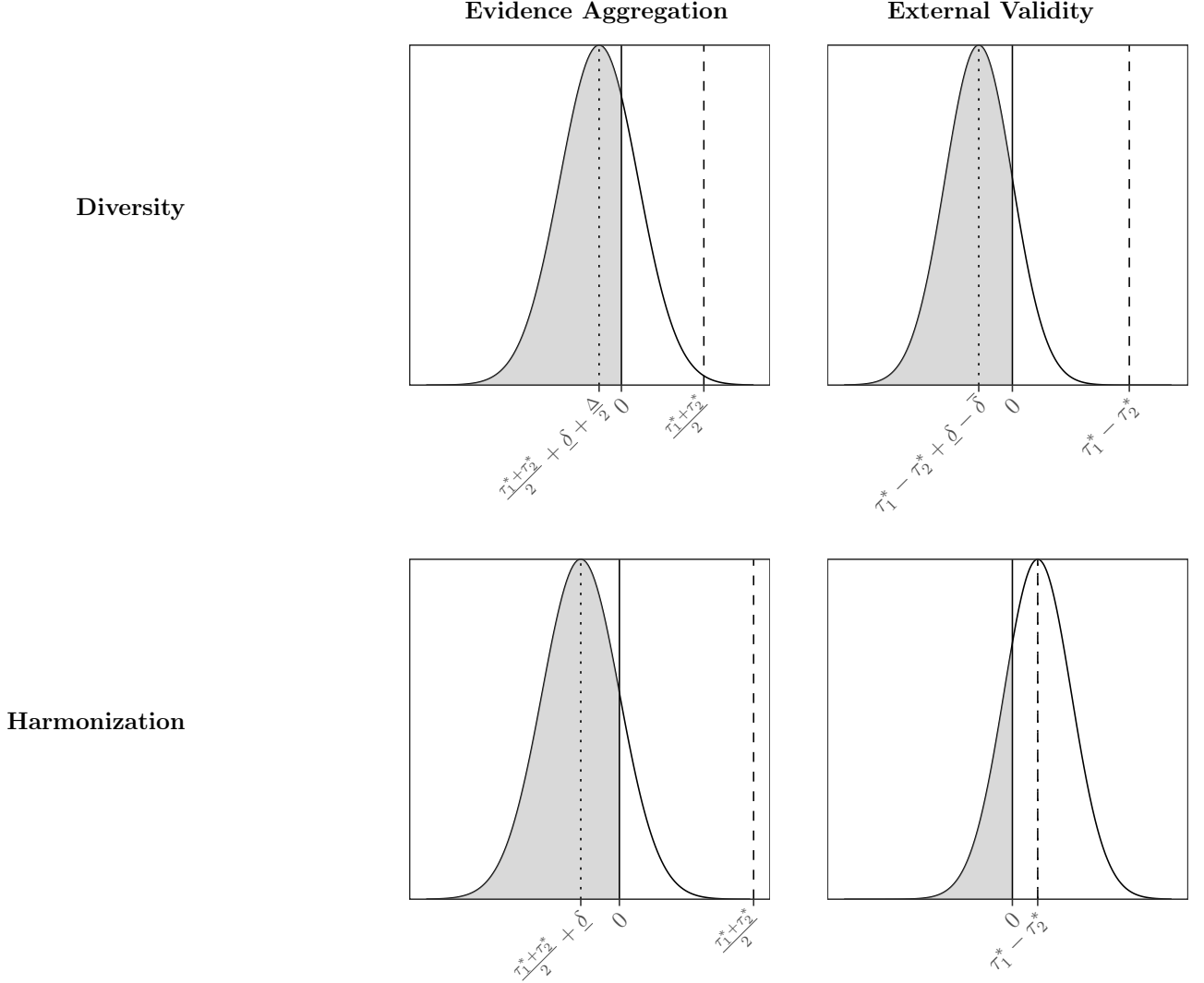


Figure 2: Worst-case distribution of estimates by estimand and harmonization choice

Figures display the probability density function $f(\hat{\tau})$ of the estimates for treatment effects and design artifacts that maximize the decision-maker's regret. Dashed vertical lines show estimand values, in terms of τ_1^*, τ_2^* that maximize regret, though τ_1^* and τ_2^* are not unique. Different combinations of context-level effects can produce the same average or difference. We focus on positive estimands rather than symmetric maxima in the negative domain. Dotted vertical lines show the expectation of the estimate distribution which is shifted away from the estimand by worst-case design artifacts, i.e., $f(\delta^{\min}(h))$. All estimands are positive, i.e., the decision-maker's optimal decision is $a^* = 1$. Grey areas represent the probability that the decision-maker makes the *wrong* choice, i.e., $a^* = 0$, because her estimate $f(\hat{\tau})$ remains below zero. If the decision-maker estimates the cross-context average effect (column 1), the expectation of the estimate distribution (dotted line) is closer to the truth (dashed line) and hence the decision-maker is less likely to make the wrong choice under diversity (top) than under harmonization (bottom). If the decision-maker seeks to estimate the cross-context difference in effects (column 2), the expectation of the estimate distribution (dotted line) is far from the truth (dashed line) under diversity (top) but equals the truth under harmonization (bottom). Hence, the decision-maker is less likely to make the wrong choice under harmonization (bottom) than under diversity (top).

The worst-case regret for a given harmonization choice arises in one of two cases: either the estimand is positive and design artifacts shift estimates as far downward as possible, lowering the probability that the decision-maker correctly chooses $a = 1$; or the estimand is negative and design artifacts shift estimates as far upward as possible, increasing the likelihood that she incorrectly chooses $a = 1$.

It is not possible to explicitly solve for the vector of treatment effects $\boldsymbol{\tau}$ that maximizes the decision-maker’s regret. Although a solution exists, it need not be unique. To build intuition about the worst-case scenario, consider the case without design artifacts, so that the estimate distribution is centered on the estimand $f(\boldsymbol{\tau})$ (bottom right panel of Figure 2). As $f(\boldsymbol{\tau})$ moves away from zero, the regret from a wrong decision increases, but the probability of making such a mistake decreases. Worst-case regret therefore arises where these forces exactly offset each other.

Importantly, an explicit characterization of the worst-case $\boldsymbol{\tau}$ is unnecessary. Lemma 1 shows that a harmonization choice h is preferred if its smallest possible shift in the estimate distribution is larger (less negative) and its largest possible shift is smaller than those induced by the alternative. Intuitively, if harmonization or diversity yields estimates with uniformly smaller worst-case bias in both directions, it dominates the alternative for all values of the estimand, including the least favorable one (see Figures A2 and A3).

4.1 Evidence aggregation

Consider a decision-maker whose estimand is the cross-context average effect, $f(\boldsymbol{\tau}) = \frac{\tau_1 + \tau_2}{2}$. Under harmonization, she uses design 1 in both studies. Hence, the worst-case regret obtains if the true average effect is negative and design 1 induces the largest possible artifact $\bar{\delta}$ in both studies, or if the true average effect is positive and design 1 induces the the smallest possible artifact $\underline{\delta}$ in both studies. Equivalently, $f(\boldsymbol{\delta}^{\min}(1)) = \underline{\delta}$ and $f(\boldsymbol{\delta}^{\max}(1)) = \bar{\delta}$.

If the decision-maker diversifies, she uses design 1 in context 1 and design 2 in context 2. The resulting shift in the estimate distribution therefore depends on both artifacts, δ^1

and δ^2 . Both share the same upper bound $\bar{\delta}$ and lower bound $\underline{\delta}$. Yet, we have assumed that $|\delta_1 - \delta_2| \geq \Delta > 0$. Hence, the maximal shift of the estimate distribution under diversity is

$$f(\boldsymbol{\delta}^{\max}(0)) = \frac{\bar{\delta} + \bar{\delta} - \Delta}{2} = \bar{\delta} - \frac{\Delta}{2}$$

and the minimal shift is

$$f(\boldsymbol{\delta}^{\min}(0)) = \frac{\underline{\delta} + \underline{\delta} + \Delta}{2} = \underline{\delta} + \frac{\Delta}{2}.$$

Since the largest possible shift of the estimate distribution is larger and the smallest is smaller under design harmonization than under diversity, harmonization generates higher regret for the decision-maker. Figure 2 highlights a subtle point: the cross-context average effect that is least favorable for the decision-maker differs under harmonization and diversity. Nevertheless, the decision-maker strictly prefers design diversity, since for any value of the average effect – including the least favorable one – harmonization makes an incorrect choice more likely than diversity.

Proposition 1. *For $f(\boldsymbol{\tau}) = \frac{\tau_1 + \tau_2}{2}$, $R(1) > R(0)$ and hence $h^* = 0$, i.e., the decision-maker prefers research design diversity.*

Intuitively, design diversity protects the decision-maker from worst-case scenarios in which a single design induces large artifacts across all studies. As long as we are willing to assume that different measurement strategies do not generate identical artifacts, measurement diversity improves what the decision-maker can learn about the cross-study average effect in the worst case.

4.2 External validity

Next, consider a decision-maker whose estimand is the cross-context difference in effects, $f(\boldsymbol{\tau}) = \tau_1 - \tau_2$. The advantage of design harmonization is immediately apparent: when artifacts are constant across studies, they difference out, yielding valid estimates of the

cross-context difference in treatment effects, i.e., $f(\boldsymbol{\delta}^{\min}(1)) = f(\boldsymbol{\delta}^{\max}(1)) = 0$. Under design diversity, by contrast, the location of the estimate distribution reflects both true cross-context differences and differences in design artifacts. Here, the worst case arises when one artifact takes its maximal value and the other its minimal value, so that $f(\boldsymbol{\delta}^{\min}(0)) = \underline{\delta} - \bar{\delta}$ and $f(\boldsymbol{\delta}^{\max}(0)) = \bar{\delta} - \underline{\delta}$. As illustrated in Figure 2, harmonization is preferred.

Proposition 2. *For $f(\boldsymbol{\tau}) = \tau_1 - \tau_2$, $R(0) > R(1)$ and hence $h^* = 1$, i.e., the decision-maker prefers research design harmonization.*

If artifacts are context-specific, they may not fully cancel even under harmonization. Below, we allow for imperfect cross-context dependence in design artifacts. We show that harmonization remains optimal for assessing external validity as long as artifacts are not completely independent across contexts.

5 Robustness of results

5.1 Optimal decision-rules

We have assumed the decision-maker acts according to the decision-rule given in equation 4. However, one may wonder whether our conclusions still hold if the decision-maker could adjust her decision-rule based on her harmonization choice. Here, we suppose that the decision-maker chooses $a = 1$ whenever her estimate $f(\hat{\boldsymbol{\tau}})$ exceeds some cutoff $c \in \mathbb{R}$ that may be different from zero and allow her to choose the optimal cutoff in a minimax sense, i.e., the decision-maker solves the following problem

$$\min_{h,c} R(h,c), \text{ where } R(h,c) := \max_{\boldsymbol{\tau}, \boldsymbol{\delta}(h)} r(\boldsymbol{\tau}, \boldsymbol{\delta}(h), c). \quad (8)$$

Equation 4 turns out to be the optimal decision-rule if the decision-maker's estimand is the cross-context difference in treatment effects. If she seeks to estimate the cross-context average,

the decision-maker can guarantee herself a smaller worst-case regret by setting the cutoff equal to the midpoint between the upper and the lower bound of the two design artifacts.

Proposition 3. *The decision-maker’s optimal cutoff c^* is*

$$c^* = \begin{cases} \frac{\delta + \bar{\delta}}{2} & \text{if } f(\mathbf{x}) = \frac{x_1 + x_2}{2} \\ 0 & \text{if } f(\mathbf{x}) = x_1 - x_2. \end{cases}$$

In neither case does the decision-maker benefit from adjusting her cutoff based on her harmonization choice. As a result, the relative attractiveness of research design harmonization and diversity remains unaffected by the decision-maker’s optimal choice of c . Real-world researchers are of course unlikely to be able to optimize their decision-rule in accordance with proposition 3, because they will not know the bounds on the design artifacts. The result is nonetheless important since it shows that our conclusions survive if we allow the decision-maker to make the best theoretically possible use of the data.

5.2 Joint tests rather than direct aggregation

We have focused on the cross-context average as a way to aggregate evidence, because this estimand is common in empirical practice, but of course other ways to aggregate evidence exist. One possibility is to jointly test a set of null hypotheses, each pertaining to a context-level effect τ_i . This approach can be helpful if effect estimates are incommensurable, e.g., because they are located on different scales that cannot be easily compared. In appendix D, we consider a decision-maker who conducts an intersection-union test of a global null hypothesis that consists of the intersection of a set of component null hypotheses against the alternative hypothesis that at least one of the component null hypotheses is false (Berger, 1997). We show that research design diversity remains the preferred choice for this alternative form of evidence aggregation. We also explain the relationship to the partial conjunction test of Egami and Hartman (2023).

5.3 A Bayesian framework

Our minimax regret analysis avoids many parametric assumptions including the stipulation of subjective prior distributions. Yet, one may worry that this analysis is prone to favor research design diversity because our decision-maker focuses on the worst case scenario. In appendix E, we derive the same results from a Bayesian setup. We consider a parametric model that resembles a random effects meta-analysis (e.g., Gelman et al., 2013, chap. 5.6). As before, design diversity is optimal for evidence aggregation estimands, whereas harmonization dominates for external validity assessments.

5.4 Imperfect dependence of artifacts under harmonization

Core to our notion of design harmonization is that it introduces dependence in design artifacts across studies. Yet, to generate our results, we do not need this dependence to be perfect. For example, suppose design j introduces an artifact δ_i^j in context i that is the weighted sum of the context-invariant component δ^j and a context-specific component $\eta_i^j \in \mathbb{R}$:

$$\delta_i^j = \omega\delta^j + (1 - \omega)\eta_i^j,$$

where $\omega \in [0, 1]$ parameterizes the strength of the cross-context dependence in design artifacts. If $\omega = 1$, design artifacts of a given design are constant across contexts as in our main analysis. If $\omega = 0$, design artifacts of a given design are entirely context-specific. As before, we assume the range of artifacts that each design could induce in a given context is the same across designs such that the decision-maker has no a priori reason to prefer one design over the other. Specifically, we maintain our previous assumptions on δ^j and also assume that $\eta_i^j \in [\underline{\eta}_i, \bar{\eta}_i]$ for $j \in \{1, 2\}$ and $i \in \{1, 2\}$, with $\underline{\eta}_i, \bar{\eta}_i \in \mathbb{R}$ and $\underline{\eta}_i < \bar{\eta}_i$. Hence, the plausible range of design artifact δ_i^j for $j \in \{1, 2\}$ in context i is $[\underline{\delta}_i, \bar{\delta}_i] = [\omega\underline{\delta} + (1 - \omega)\underline{\eta}_i, \omega\bar{\delta} + (1 - \omega)\bar{\eta}_i]$. Note that although we assume ambiguity-equivalence within contexts, the plausible range of artifacts may differ across contexts for both designs. For example, experimenter demand may be more

severe in some contexts for all outcome measures.

Proposition 4. For $f(\boldsymbol{\tau}) = \frac{\tau_1 + \tau_2}{2}$, $R(1) \geq R(0)$. For $f(\boldsymbol{\tau}) = \tau_1 - \tau_2$, $R(1) \leq R(0)$. These inequalities are strict if and only if $\omega > 0$.

As long as harmonization introduces *any* dependence in design artifacts across contexts, the decision-maker prefers design diversity for evidence aggregation and harmonization for external validity assessments. Only when artifacts are completely independent across contexts ($\omega = 0$) does the decision-maker become indifferent. In this case, harmonization is essentially meaningless. This indifference of course reflects our assumption that harmonization is costless: if harmonization introduces costs, say, in terms of coordination across researchers, design diversity would be preferred for both estimands. Figure 3 shows how the absolute difference in regret between harmonization and diversity – and thus the stakes of the choice between them – increases with cross-context dependence in design artifacts. Appendix F shows these results generalize to our Bayesian analysis.

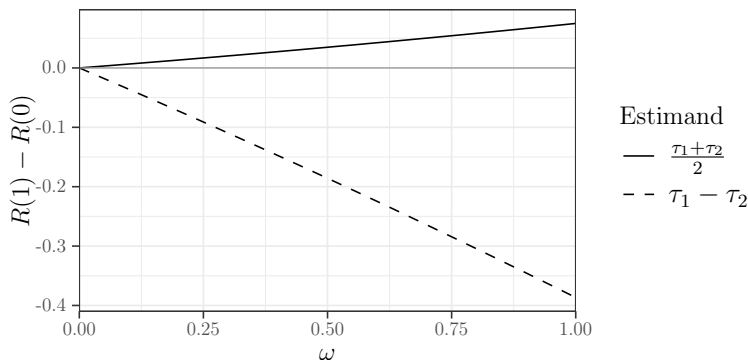


Figure 3: The difference between the maximum regret under design harmonization and diversity as a function of the cross-context dependence in design artifacts under harmonization $R(1)$ is maximum regret under harmonization, $R(0)$ is maximum regret under diversity. The decision-maker seeks to minimize her maximum regret. $\sigma^2 = 2$, $\bar{\delta} = 0.8$, $\underline{\delta} = -0.5$, $\eta_1 = -0.1$, $\bar{\eta}_1 = 0.25$, $\eta_2 = -0.2$, $\bar{\eta}_2 = 0.35$, $\Delta = 0.5$.

5.5 More than two contexts

Our main analysis considers two contexts, but similar logics arise with $N > 2$ contexts, where we assume for simplicity that N is even. $\boldsymbol{\tau}$ now is an N -dimensional vector of treatment

effects, one for each context. To aggregate evidence, the decision-maker estimates the average effect $f(\boldsymbol{\tau}) = \frac{\sum_{i=1}^N \tau_i}{N}$. For external validity assessments, we consider two estimands. The first is the difference between the effect in some context i and the average effect across all other $N - 1$ contexts, $f(\boldsymbol{\tau}) = \tau_i - \frac{\sum_{i' \neq i}^N \tau_{i'}}{N-1}$. The second is the difference between the average effect across two subgroups $p \in \{X, Y\}$, for example, democracies and autocracies. W.l.o.g., let subgroup X comprise contexts $1, \dots, n_X$ and subgroup Y the remaining $N - n_X$ contexts, so that $f(\boldsymbol{\tau}) = \frac{\sum_{i=1}^{n_X} \tau_i}{n_X} - \frac{\sum_{i=n_X+1}^N \tau_i}{N-n_X}$. As before, harmonization assigns design 1 to all contexts. Under diversity, the decision-maker uses design 1 in $n^{(1)} < N$ contexts and design 2 in all other $N - n^{(1)}$ contexts. Let ξ_p denote the share of contexts in group p assigned design 1 under diversity. We return to our base setup where design j introduces the same artifact δ^j across contexts.

Proposition 5. *The decision-maker's optimal choice h^* depends on her estimand as follows:*

- For $f(\boldsymbol{\tau}) = \frac{\sum_{i=1}^N \tau_i}{N}$, $R(1) > R(0)$, and hence $h^* = 0$. $R(0)$ is minimal for $n^{(1)*} = \frac{N}{2}$.
- For $f(\boldsymbol{\tau}) = \tau_i - \frac{\sum_{i' \neq i}^N \tau_{i'}}{N-1}$, $R(1) < R(0)$, and hence $h^* = 1$.
- For $f(\boldsymbol{\tau}) = \frac{\sum_{i=1}^{n_X} \tau_i}{n_X} - \frac{\sum_{i=n_X+1}^N \tau_i}{N-n_X}$, $R(1) = R(0)$ and hence $h^* \in \{0, 1\}$ if and only if $\xi_X = \xi_Y = \frac{n^{(1)}}{N}$.⁷ Otherwise $R(1) < R(0)$, and hence $h^* = 1$.

As in the two-context case, design diversity is preferred for evidence aggregation because it guards against the worst case in which a single biased design is used in all contexts. Moreover, the decision-maker minimizes maximal regret by using each design in exactly half of the contexts. If one design were used more heavily, the worst case would arise when that design introduced the larger artifact. The decision-maker therefore has incentives to reduce reliance on the more heavily used design until both designs are used equally, that is, to diversify as much as possible.

⁷This condition requires that n_X and n_Y are divisible by $\frac{N}{n^{(1)}}$, a mild requirement given that the decision-maker can choose $n^{(1)}$.

For external validity assessments defined as the difference between the effect in a single context and the average effect across a group of contexts, harmonization is preferred, as it ensures that artifacts cancel out. When the estimand is the difference between average effects across two subsets of contexts, however, the decision-maker has more flexibility: artifacts cancel out either under complete harmonization or under design diversity that uses each design in the same proportion within both groups. For example, the decision-maker performs equally well under complete harmonization and when assigning design 1 to one third and design 2 to two thirds of the contexts in each group. The latter approach is preferable if the decision-maker is interested in *both* evidence aggregation and external validity assessments.

Considering averages across groups of contexts can therefore soften the design trade-offs faced by decision-makers and shows that some design diversity can be desirable – even for external validity assessments. That said, the harmonization motif does not disappear. To see this, imagine a decision-maker has access not only to two but to as many ambiguity-equivalent designs as contexts. A decision-maker concerned only with evidence aggregation would use a different design in each context, whereas one interested in differences in average effects across subgroups would repeat the same design across at least some contexts in both groups.

6 Extensions

6.1 Research design as sampling variance

So far, we have conceptualized research designs as introducing design artifacts. Alternatively, research designs may affect the spread of the estimate distribution. The available outcome measures may accurately capture the outcome of interest *on average* but some measures may be noisier than others. Treatment variations may have differential consequences for the variability of the estimate distribution as well. For example, the standard error of the difference-in-means estimator increases with the covariance between treated and untreated potential outcomes (Gerber and Green, 2012, p. 57). Inter-group contact interventions that

are more effective at increasing tolerance among already tolerant individuals will thus produce noisier estimates than inter-group contact interventions that unfold their greatest effects among the a priori least tolerant individuals. Such differences may arise from incidental features of the intervention – like participants’ freedom to avoid contact.

Consider a version of our framework in which the variance σ^2 of the statistical noise term ϵ_i is specific to research design j , i.e., $\epsilon_i \sim \mathcal{N}(0, \sigma^{(j)2})$, but design artifacts are absent, i.e., $\delta^j = 0$ for $j \in \{1, 2\}$. As before, the decision-maker faces ambiguity over the quality of research designs, but quality now hinges on differences in precision. We presume that $\sigma^{(j)2} \in [\underline{\sigma}^2, \bar{\sigma}^2]$ for $j \in \{1, 2\}$ with $\underline{\sigma}^2 > 0$, $\bar{\sigma}^2 > 0$, and $\underline{\sigma}^2 < \bar{\sigma}^2$, i.e., the variance of the estimate distribution induced by each design lies within some plausible range and this range is identical across designs. Hence, designs remain *ambiguity-equivalent*. We also assume $|\sigma^{(1)2} - \sigma^{(2)2}| \geq \Delta_\sigma > 0$. For simplicity, we here return to the two-context case, though our results generalize directly to the case with more than two contexts.

In this framework, the decision-maker always prefers design diversity. Intuitively, the decision-maker’s regret increases in the variability of her estimates. Hence, the decision-maker seeks to minimize the largest possible variance. Irrespective of the estimand, harmonization risks a worst case scenario in which the decision-maker’s chosen design produces estimates that are highly imprecise in all studies. Diversification counters this risk.

Proposition 6. *If research designs differ only in terms of their variance $\delta^{(j)2}$ and $\delta^1 = \delta^2 = 0$, $R(1) > R(0)$ and hence $h^* = 0$, i.e., the decision-maker prefers research design diversity.*

In appendix G, we allow designs to introduce artifacts *and* change the variance of the estimate distribution. In this case, the decision-maker may prefer a more variable over a precise design if the design is also severely biased. Nonetheless, our conclusions remain unchanged – design diversity clearly dominates in the evidence aggregation case, while the optimal choice for external validity assessments depends on the relative contribution of worst-case design artifacts and worst-case variances to the decision-maker’s maximum regret.

6.2 Within-study design diversity

Appendix H studies design diversity within a single study, using a simple example of outcome measures. The exercise could be easily repeated for within-study diversity in treatment operationalizations (i.e., “stimulus sampling,” Wells and Windschitl, 1999; Monin and Oppenheimer, 2014). We consider a decision-maker who runs a single experiment and chooses between using the same measure for all units in the experiment (harmonization) or randomly assigning some share of units to one measure and the rest to the other (diversity). The decision-maker cannot use both outcome measures for all units, e.g., because of a budget constraint. Outcome measures may introduce systematic measurement error and the decision-maker faces equivalent ambiguity over these errors.

We show that the choice between within-study harmonization and diversity can involve a bias-variance trade-off. Design diversity is beneficial for the estimation of average treatment effects, because it guards against the worst-case scenario of using a single highly biased outcome measure for all units. However, design diversity also makes potential outcomes more variable and hence increases the variance of the estimate distribution. Which concern dominates depends on the constellation of potential outcomes and worst-case biases.

However, the decision-maker can eliminate the impact of design diversity on the variance of her estimates by randomly assigning units to one of the two outcome measures and then assigning units to treatment *within blocks* that reflect the outcome measure to which units were assigned. Since units within the same block use the same outcome measure, estimates under research design diversity with this blocking structure are no more variable than estimates under harmonization. Thus, the choice between within-study harmonization and diversity hinges on bias-concerns only and on this dimension, design diversity minimizes the decision-maker’s maximum regret.

7 Practical lessons

Our framework yields concrete lessons for designing collaborative studies or meta-analyses.

1. *Do not default to research design harmonization.* Whether harmonization or design diversity is more appropriate depends on inferential goals and concerns. Of course, researchers should select treatment and outcome operationalizations that plausibly capture underlying theoretical constructs. However, even after such conceptual harmonization, multiple viable operationalizations typically remain. Sometimes the choice among them is straightforward because one operationalization is clearly superior. Or, harmonization concerns may be moot if all treatment variations and outcome measures can be implemented within each study. Often, however, budgetary and feasibility constraints – combined with weak prior beliefs about the merits of alternative operationalizations – leave researchers choosing among several designs of plausibly similar quality. By our analysis, researchers should pursue harmonization – typically a costly endeavor (Blair and McClendon, 2021) – *only* under the following conditions:

- Researchers’ primary objective are external validity assessments, i.e., the estimation of cross-context differences in effects.
- The primary inferential concern regarding outcome and treatment operationalizations is that they may introduce artifacts – i.e., bias effect estimates – and not just noise.
- It is plausible that using the same operationalizations across contexts leads to design artifacts that are more similar than using different operationalizations.

Otherwise, and particularly in the scenario in which researchers seek to accumulate evidence with a cross-context average or joint null hypothesis test, design diversity increases robustness. Even if the above conditions are met, researchers interested in comparing average effects across two sub-groups of contexts can diversify within each sub-group.

2. *Partially harmonize for multiple inferential targets.* Many studies will seek both to estimate average effects *and* assess whether effects vary across contexts. For a meta-analysis, one could optimize for both goals by subsetting on studies with similar designs when

assessing cross-context external validity but including diverse studies in assessing aggregate effects. When designing prospective cross-context studies, one could implement two treatment variations and collect two outcome measures in each context. Researchers can then harmonize one treatment and one outcome measure across studies but allow for variation in the other. Harmonized design elements would be used to assess effect heterogeneity, while estimation of average effects would use of the full set of operationalizations.

We now discuss how these lessons apply to two recent cross-context studies. Both are exemplary in terms of rigor, innovation, and breadth, but our analysis offers guidance on increasing their inferential value.

7.1 The generalizability of IR experiments

Bassan-Nygate et al. (2025) replicate four prominent IR survey experiments originally conducted in the United States across seven democracies – Brazil, Germany, India, Israel, Japan, Nigeria, and the United States – introducing purposeful variation in contexts. The authors’ primary goal is evidence aggregation: they assess the number of countries in which the treatment effect has a particular sign through a partial conjunction test and estimate the cross-context average effect through a meta-analysis. The survey experiments are harmonized across all seven contexts, and, for the most part, use the same treatment vignettes and outcome measures as the original US-based experiment (with exceptions detailed in their Table A3). The authors also harmonized several ancillary design features: all survey experiments were conducted by the same survey firm, and fielded at the same time. The authors find that pre-existing findings from the United States largely generalize.

We use the survey experiment testing the democratic peace theory to consider possible design artifacts and highlight how design diversity could be beneficial given the study’s inferential objective. This exercise is intended as an illustration rather than as a criticism. The democratic peace theory predicts that citizens in democracies are less willing to go to war with other democracies, possibly because of citizens’ perceptions of threat, the costs

and likelihood of success of war, and moral considerations. The survey experiment presents respondents with a hypothetical scenario in which a country is developing nuclear weapons, randomly varying whether the country is described as a democracy or an autocracy. The outcome measure is a survey item eliciting respondents' support for attacking the country's nuclear sites.⁸ Note that the democratic peace theory is not specific about the cause of war being nuclear weapons. The focus on nuclear weapons in the treatment vignette is thus an operationalization choice and not a consequence of the theory.

Survey experiments can produce potentially confounding design artifacts when changes to one piece of information affect respondents' beliefs about background conditions – so-called “information leakage” (Dafoe, Zhang and Caughey, 2018). To their credit, Bassan-Nygate et al. (2025) examine this issue and rule out information leakage as a reason for the null effect in the India sample. Yet, information leakage could affect conclusions beyond this context. In their appendix the authors note “some evidence that respondents in the non-Democracy condition were more likely to report that they thought of a specific country.” For example, citizens assigned to the autocracy condition may tend to interpret the country in the vignette to be North Korea, which could raise support for war for reasons other than autocracy.

One way to address these concerns would be to vary the operationalizations of both the treatment and the outcomes – for example, by using a vignette in which a country threatens to invade a neighbor. Such alternative operationalizations may also suffer from information leakage, but it does not seem self-evident that they are better or worse. Reliance on a variety of operationalizations would guard against a scenario in which the estimates in all contexts are confounded by a shared design artifact. Introducing diversity in vignettes would of course complicate comparisons of the new estimates to those from the pre-existing US experiments. Nonetheless, it would advance the authors' inferential objective of providing robust aggregate evidence on democratic peace theory. The same reasoning extends to ancillary design features beyond treatment and outcome operationalizations, such as the choice of survey firm or the

⁸Bassan-Nygate et al. (2025) add an additional outcome because they worry about floor effects.

timing of fieldwork.

7.2 Community policing in the Global South

The Metaketa initiative of the Evidence in Governance and Politics (EGAP) network organizes coordinated experiments to study common research questions across contexts. The fourth Metaketa round consisted of six coordinated field experiments on “community policing,” each implemented by a different research team, examining effects on citizen trust in police and crime in Brazil, Colombia, Liberia, Pakistan, the Philippines, and Uganda (Blair et al., 2021). The project used a random effects meta analysis to estimate the cross-context average effect (evidence aggregation). The authors allowed for variation in treatment operationalizations across contexts, but they harmonized the crime and citizen trust outcome measures, with minor contextual adaptations. The preregistered meta-analysis found that community policing interventions did not increase citizen trust in the police or reduce crime.

Allowing for diversity in treatment operationalizations is well aligned with the evidence aggregation objective. The general conclusion that community policing is ineffective is more compelling because it is based interventions that vary in duration, training, and modes of citizen participation. That said, diversity in treatment operationalizations makes it difficult to compare effects across study contexts. A feature of the Metaketa initiative is that each individual trial features a second treatment arm. Most researchers used the second arm to target a different conceptual treatment. While this approach aligns with researcher incentives to seek novelty, our analysis highlights that having each context implement one common and one *alternative operationalization of the same conceptual treatment* would be helpful for studying *both* cross-context aggregate and differences in effects.

The harmonization of outcome measures detracts from the goal of evidence aggregation. The authors’ measure of citizens’ trust in police, for example, was an index of two items with identical wording (up to translation) and response options across contexts. Artifacts are conceivable: the items have a repetitive response format that may result in respondents just

repeating the same answer with little consideration, and they have a middle category that allows respondents an “easy way out” from giving careful consideration. Survey items also raise concerns about social desirability biases. One could certainly imagine phrasing survey questions differently, or using other measures like questions about concrete police services or behavioral measures. Any outcome operationalization has its own problems, but diversity would increase robustness to a scenario where results are an artifact of the specific measures that the authors chose to implement in all contexts. This concern may seem small here since each individual study – while harmonized – includes a large variety of outcome measures, reducing robustness concerns. However, harmonization also came with high costs in time and effort. Given that harmonization does not further the study’s primary inferential objective, a sensible alternative would be to ensure that all relevant conceptual outcomes are measured in all contexts while allowing for diversity in measurement protocols.

8 Conclusion

This paper offers a framework for research design in multi-context studies, clarifying the trade-offs between harmonization and diversification. While harmonization enhances the ability to detect effect heterogeneity by controlling for design-induced artifacts, diversity mitigates the risk of systematic biases when aggregating evidence across studies. Design decisions should be aligned with the primary inferential objective – be it to estimate a generalizable average effect or to test for cross-context variation. Our findings caution against defaulting to uniformity in multi-site research and call attention to the epistemic value of purposeful design variation.

We have interpreted research designs in terms of outcome and treatment operationalizations. But any study characteristic that can cause study estimates to systematically differ from the quantity of interest could be a design element in our main framework. This includes implementation details and could even extend to identification or sampling strategies, insofar

as “ambiguity equivalence” is plausible.

Several avenues for further research stand out. First, we separate questions of design harmonization from research design quality by assuming a choice over ambiguity-equivalent designs. Future work could explicitly model the trade-offs that arise in the presence of both quality differences and variation in learning objectives. Second, future research could consider additional estimands. For example, we do not consider how design choices impact the decision-maker’s ability to learn about design artifacts (Slough and Tyson, 2024) or unbundling the ingredients through which a treatment produces its effects (Callis, Dunning and Tuñón, 2023). Third, we do not explicitly address research timing or dynamics. Future work could examine how the choice between design harmonization and diversity is shaped by the accumulation of evidence over time. Finally, we model a decision-maker who is purely learning-motivated. One could extend to career-concerned researchers interacting strategically.

References

- Abadie, Alberto. 2020. “Statistical Nonsignificance in Empirical Economics.” *American Economic Review: Insights* 2(2):193–208.
- Abramson, Scott F, Korhan Koçak and Asya Magazinnik. 2022. “What do we learn about voter preferences from conjoint experiments?” *American Journal of Political Science* 66(4):1008–1020.
- Allport, Gordon W. 1954. *The Nature of Prejudice*. Garden City, NJ: Anchor.
- Azevedo, Eduardo M, Alex Deng, José Luis Montiel Olea, Justin Rao and E Glen Weyl. 2020. “A/b testing with fat tails.” *Journal of Political Economy* 128(12):4614–000.
- Banerjee, Abhijit, Dean Karlan and Jonathan Zinman. 2015. “Six randomized evaluations of microcredit: Introduction and further steps.” *American Economic Journal: Applied Economics* 7(1):1–21.
- Banerjee, Abhijit V, Sylvain Chassang and Erik Snowberg. 2017. Decision theoretic approaches to experiment design and external validity. In *Handbook of Economic Field Experiments*. Vol. 1 Elsevier pp. 141–174.
- Banerjee, Abhijit V, Sylvain Chassang, Sergio Montero and Erik Snowberg. 2020. “A theory of experimenters: Robustness, randomization, and balance.” *American Economic Review* 110(4):1206–30.
- Bassan-Nygate, Lotem, Jonathan Renshon, Jessica LP Weeks and Chagai M Weiss. 2025. “The generalizability of IR experiments beyond the United States.” *American Political Science Review* 119(4):1649–1664.
- Berger, Roger L. 1997. Likelihood ratio tests and intersection-union tests. In *Advances in statistical decision theory and applications*. Springer pp. 225–237.

- Blair, Graeme and Gwyneth McClendon. 2021. Conducting Experiments in Multiple Contexts. In *Advances in Experimental Political Science*, ed. James N. Druckman and Donald P. Green. Cambridge: Cambridge University Press pp. 411 – 428.
- Blair, Graeme, Jasper Cooper, Alexander Coppock and Macartan Humphreys. 2019. “Declaring and Diagnosing Research Designs.” *American Political Science Review* 113(3):838–859.
- Blair, Graeme, Jeremy M Weinstein, Fotini Christia, Eric Arias, Emile Badran, Robert A Blair, Ali Cheema, Ahsan Farooqui, Thiemo Fetzer, Guy Grossman et al. 2021. “Community policing does not build citizen trust in police or reduce crime in the Global South.” *Science* 374(6571):eabd3446.
- Callis, Anna, Thad Dunning and Guadalupe Tuñón. 2023. Knowledge Accumulation through Natural Experiments. In *Oxford Handbook of Engaged Methodological Pluralism in Political Science*, ed. Janet M. Box-Steffensmeier, Dino P. Christenson and Valeria Sinclair-Chapman. Oxford: Oxford University Press.
- Campbell, Donald T., Julian C. Stanley and N. L. Gage. 1966. *Experimental and quasi-experimental designs for research*. Chicago: R. McNally.
- 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.
- Cronbach, Lee J and Paul E Meehl. 1955. “Construct Validity in Psychological Tests.” *Psychological Bulletin* 52(4):281.
- Dafoe, Allan, Baobao Zhang and Devin Caughey. 2018. “Information equivalence in survey experiments.” *Political Analysis* 26(4):399–416.
- Danchev, Valentin, Andrey Rzhetsky and James A Evans. 2019. “Meta-Research: Centralized scientific communities are less likely to generate replicable results.” *Elife* 8:e43094.

- Deaton, Angus. 2010. “Instruments, randomization, and learning about development.” *Journal of economic literature* 48(2):424–55.
- Deaton, Angus and Nancy Cartwright. 2018. “Understanding and misunderstanding randomized controlled trials.” *Social Science & Medicine* 210:2–21.
- Dominitz, Jeff and Charles Manski. 2017. “More data or better data? A statistical decision problem.” *The Review of Economic Studies* 84(4):1583–1605.
- Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D Hyde, Craig McIntosh and Gareth Nellis. 2019. *Information, accountability, and cumulative learning: Lessons from Metaketa I*. Cambridge University Press.
- Egami, Naoki and Diana Da In Lee. 2024. “Designing Multi-Site Studies for External Validity: Site Selection via Synthetic Purposive Sampling.” *Available at SSRN 4717330* .
- Egami, Naoki and Erin Hartman. 2020. “Elements of External Validity: Framework, Design, and Analysis.” *Design, and Analysis (June 30, 2020)* .
- Egami, Naoki and Erin Hartman. 2023. “Elements of external validity: Framework, design, and analysis.” *American Political Science Review* 117(3):1070–1088.
- Frankel, Alexander and Maximilian Kasy. 2022. “Which Findings Should Be Published?” *American Economic Journal: Microeconomics* 14(1):1–38.
- Fu, Jiawei and Donald P Green. 2025. “Causal Inference for Experiments with Latent Outcomes: Key Results and Their Implications for Design and Analysis.” *arXiv preprint arXiv:2505.21909* .
- Galos, Diana Roxana and Alexander Coppock. 2023. “Gender composition predicts gender bias: A meta-reanalysis of hiring discrimination audit experiments.” *Science Advances* 9(18):eade7979.

- Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari and Donald B Rubin. 2013. *Bayesian data analysis*. CRC press.
- Gerber, Alan S and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation*. New York: WW Norton.
- Gerring, John. 2023. The Cumulation Problem in Political Science: Toward Standardization. In *The Oxford Handbook of Engaged Methodological Pluralism in Political Science, Vol. 1*, ed. Janet M Box-Steffensmeier, Dino Christenson and Valeria Sinclair-Chapman. Oxford University Press.
- Green, Donald P and Alan S Gerber. 2019. *Get Out The Vote: How to Increase Voter Turnout*. 4 ed. Washington, D.C.: Brookings Institution Press.
- Humphreys, Macartan and Alan M Jacobs. 2023. *Integrating Inferences: Causal Models for Qualitative and Mixed-Method Research*. Cambridge University Press.
- 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.
- Imbens, Guido W. 2010. “Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009).” *Journal of Economic literature* 48(2):399–423.
- Izzo, Federica, Torun Dewan and Stephane Wolton. 2018. “Cumulative knowledge in the social sciences: The case of improving voters’ information.” *Available at SSRN 3239047*.
- Kaelin Jr, William G. 2017. “Common pitfalls in preclinical cancer target validation.” *Nature Reviews Cancer* 17(7):425.
- Kasy, Maximilian. 2016. “Why experimenters might not always want to randomize, and what they could do instead.” *Political Analysis* 24(3):324–338.

- Kasy, Maximilian and Jann Spiess. 2022. “Rationalizing Pre-Analysis Plans: Statistical Decisions Subject to Implementability.” *Unpublished Manuscript* .
- Kertzer, Joshua D., Jonathan Renshon and Weifang Xu. 2025. “Cross-national Survey Experiments: A Practical Guide.” *Unpublished manuscript* .
- Little, Andrew T and Thomas B Pepinsky. 2021. “Learning from biased research designs.” *The Journal of Politics* 83(2):602–616.
- Lowe, Matt. 2025. “Has Intergroup Contact Delivered.” *Typescript, University of British Columbia* .
- Lucas, Jeffrey W. 2003. “Theory-testing, generalization, and the problem of external validity.” *Sociological Theory* 21(3):236–253.
- Manski, Charles F. 2000. “Identification problems and decisions under ambiguity: Empirical analysis of treatment response and normative analysis of treatment choice.” *Journal of Econometrics* 95(2):415–442.
- Manski, Charles F. 2004. “Statistical treatment rules for heterogeneous populations.” *Econometrica* 72(4):1221–1246.
- Michelangelo, Vianello, Bahník Stephan et al. 2014. “Investigating Variation in Replicability: A ‘Many Labs’ Replication Project.” *Social Psychology* 45(3):142–52.
- Milkman, Katherine L, Dena Gromet, Hung Ho, Joseph S Kay, Timothy W Lee, Pepi Pandiloski, Yeji Park, Aneesh Rai, Max Bazerman, John Beshears et al. 2021. “Megastudies improve the impact of applied behavioural science.” *Nature* 600(7889):478–483.
- Monin, Benoît and Daniel M Oppenheimer. 2014. “The limits of direct replications and the virtues of stimulus sampling.” *Social Psychology* 45(4):299–300.
- Morton, Rebecca B and Kenneth C Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. New York: Cambridge University Press.

- Mousa, Salma. 2020. “Building social cohesion between Christians and Muslims through soccer in post-ISIS Iraq.” *Science* 369(6505):866–870.
- Nosek, Brian A and Timothy M Errington. 2017. “Reproducibility in cancer biology: Making sense of replications.” *Elife* 6:e23383.
- Nosek, Brian A and Timothy M Errington. 2020. “What is replication?” *PLoS biology* 18(3):e3000691.
- Nosek, Brian A, Tom E Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S Corker, Anna Dreber, Fiona Fidler, Joe Hilgard, Melissa Kline Struhl, Michèle B Nuijten et al. 2022. “Replicability, Robustness, and Reproducibility in Psychological Science.” *Annual review of psychology* 73(1):719–748.
- Olea, José Luis Montiel, Chen Qiu and Jörg Stoye. 2023. “Decision Theory for Treatment Choice Problems with Partial Identification.” *arXiv preprint arXiv:2312.17623* .
- Open Science Collaboration. 2015. “Estimating the reproducibility of psychological science.” *Science* 349(6251):aac4716.
- Paluck, Elizabeth Levy and Donald P. Green. 2009. “Prejudice Reduction: What Works? A Review and Assessment of Research and Practice.” *Annual Review of Psychology* 60:339–367.
- Paluck, Elizabeth Levy, Seth A Green and Donald P Green. 2019a. “The contact hypothesis re-evaluated.” *Behavioural Public Policy* 3(2):129–158.
- Paluck, Elizabeth Levy, Seth A. Green and Donald P. Green. 2019b. “The Contact Hypothesis Re-Evaluated.” *Behavioural Public Policy* 3(2):129–158.
- Paolini, Stefania, Jake Harwood and Mark Rubin. 2021. “Allport meets internet: A meta-analytical investigation of online intergroup contact and prejudice reduction.” *International Journal of Intercultural Relations* 81:131–141.

- Park, Jay JH, Ellie Siden, Michael J Zoratti, Louis Dron, Ofir Harari, Joel Singer, Richard T Lester, Kristian Thorlund and Edward J Mills. 2019. "Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols." *Trials* 20(1):1–10.
- Pearl, Judea and Elias Bareinboim. 2014. "External validity: From do-calculus to transportability across populations." *Statistical Science* 29(4):579–595.
- Pettigrew, Thomas F. and Linda R. Tropp. 2006. "A meta-analytic test of intergroup contact theory." *Journal of Personality and Social Psychology* 90(5):751–783.
- Pettigrew, Thomas F., Linda R. Tropp, Ulrich Wagner and Oliver Christ. 2011. "Recent advances in intergroup contact theory." *International Journal of Intercultural Relations* 35(3):271–280.
- Porter, Ethan and Yamil R Velez. 2022. "Placebo selection in survey experiments: An agnostic approach." *Political Analysis* 30(4):481–494.
- Recht, Benjamin. 2025. "A Bureaucratic Theory of Statistics." *Observational studies* 11(1):77.
- Rosenbaum, Paul R. 2010. "Evidence factors in observational studies." *Biometrika* 97(2):333–345.
- Samii, Cyrus. 2016. "Causal empiricism in quantitative research." *The Journal of Politics* 78(3):941–955.
- Scacco, Alexandra and Shana S Warren. 2018. "Can Social Contact Reduce Prejudice and Discrimination? Evidence from a Field Experiment in Nigeria." *American Political Science Review* 112(3):654–677.
- Schwarz, Susanne and Alexander Coppock. 2022. "What have we learned about gender from candidate choice experiments? A meta-analysis of sixty-seven factorial survey experiments." *The Journal of Politics* 84(2):655–668.

- Slough, Tara, Daniel Rubenson, Francisco Alpizar Rodriguez, María Bernedo Del Carpio, Mark T Buntaine, Darin Christensen, Alicia Cooperman, Sabrina Eisenbarth, Paul J Ferraro, Louis Graham et al. 2021. “Adoption of community monitoring improves common pool resource management across contexts.” *Proceedings of the National Academy of Sciences* 118(29).
- Slough, Tara and Scott A Tyson. 2023. “External Validity and Meta-Analysis.” *American Journal of Political Science* 67(2):440–455.
- Slough, Tara and Scott A Tyson. 2024. “Sign-congruence, external validity, and replication.” *Political Analysis* .
- Thelen, Kathleen and James Mahoney. 2015. Comparative-historical analysis in contemporary political science. In *Advances in Comparative-Historical Analysis*, ed. James Mahoney and Kathleen Thelen. Cambridge: Cambridge University Press Cambridge pp. 3–36.
- Vezzali, Loris, Miles Hewstone, Dora Capozza and Ralf Wolfer. 2018. “The extended contact hypothesis: A meta-analysis on 20 years of research.” *Personality and Social Psychology Review* 22(1):3–24.
- Wells, Gary L and Paul D Windschitl. 1999. “Stimulus sampling and social psychological experimentation.” *Personality and Social Psychology Bulletin* 25(9):1115–1125.
- Zhang, Yi, Melody Huang and Kosuke Imai. 2024. “Minimax Regret Estimation for Generalizing Heterogeneous Treatment Effects with Multisite Data.” *arXiv preprint arXiv:2412.11136* .
- Zhou, Yang-Yang and Jason Lyall. 2025. “Prolonged contact does not reshape locals’ attitudes toward migrants in wartime settings.” *American Journal of Political Science* 69(1):210–222.