

A Causal Framework for Cross-Cultural Generalizability



Dominik Deffner^{1,2,3} , Julia M. Rohrer⁴, and Richard McElreath¹

¹Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany; ²Science of Intelligence Excellence Cluster, Technical University Berlin, Berlin, Germany; ³Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany; and ⁴Department of Psychology, Leipzig University, Leipzig, Germany

Advances in Methods and Practices in Psychological Science
 July-September 2022, Vol. 5, No. 3,
 pp. 1–18
 © The Author(s) 2022
 Article reuse guidelines:
 sagepub.com/journals-permissions
 DOI: 10.1177/25152459221106366
 www.psychologicalscience.org/AMPPS



Abstract

Behavioral researchers increasingly recognize the need for more diverse samples that capture the breadth of human experience. Current attempts to establish generalizability across populations focus on threats to validity, constraints on generalization, and the accumulation of large, cross-cultural data sets. But for continued progress, we also require a framework that lets us determine which inferences *can* be drawn and how to make informative cross-cultural comparisons. We describe a generative causal-modeling framework and outline simple graphical criteria to derive analytic strategies and implied generalizations. Using both simulated and real data, we demonstrate how to project and compare estimates across populations and further show how to formally represent measurement equivalence or inequivalence across societies. We conclude with a discussion of how a formal framework for generalizability can assist researchers in designing more informative cross-cultural studies and thus provides a more solid foundation for cumulative and generalizable behavioral research.

Keywords

cross-cultural research, generalizability, WEIRD-samples problem, causal inference, poststratification, open data, open materials

Received 10/11/21; Revision accepted 4/21/22

The behavioral and social sciences have been criticized for relying excessively on WEIRD samples in which most participants are Western, educated, and from industrialized, rich, and democratic countries (Apicella et al., 2020; Henrich, 2020; Henrich et al., 2010; Muthukrishna et al., 2020). Research has established substantial cross-cultural variation in key psychological domains, such as thinking styles (e.g., Masuda & Nisbett, 2001; Nisbett & Miyamoto, 2005), economic preferences (e.g., Falk et al., 2018; Gächter & Schulz, 2016), personality structure (e.g., Smaldino et al., 2019), and moral judgments (e.g., Awad et al., 2020; Curtin et al., 2020), and furthermore demonstrated that WEIRD subjects often represent outliers among present-day societies (Apicella et al., 2020). These findings make it clear that broad, *unqualified* generalizations about human psychology based on WEIRD samples alone are rarely justified.

Fortunately, behavioral scientists increasingly acknowledge the problem. Cross-cultural psychologists and anthropologists are making progress in documenting variation in psychological phenomena (Apicella et al., 2020). In addition to long-term fieldwork and experimental comparisons across societies, large-scale collaborative projects have started compiling extensive data sets addressing cross-cultural variation and commonality in domains such as music (Mehr et al., 2019), social perception (Jones et al., 2021), and economic (Henrich et al., 2001) and moral decision-making (Awad et al., 2018). Accompanying the surge in cross-cultural studies,

Corresponding Author:

Dominik Deffner, Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany
 Email: deffner@mpib-berlin.mpg.de



researchers increasingly consider the historical and political contexts of their work as well as its ethical ramifications (e.g., Broesch et al., 2020; Clancy & Davis, 2019; Ghai, 2021; Urassa et al., 2021).

New data bring new problems. How can valid comparisons and conclusions be derived from cross-cultural samples? Just as there are many ways to misinterpret data from a single society, there are even more ways to misinterpret differences or similarities between societies. In each case, one must first generalize from each sample to each population before valid comparisons can be made between populations. This is a generalizability problem on a global scale.

Methodologists have long discussed the importance of generalizability or “external validity” and its relationship to other kinds of validity (internal, statistical conclusion and construct validity; e.g., Berkowitz & Donnerstein, 1982; Calder et al., 1983; Campbell, 1957; Cock & Campbell, 1976; Winer, 1999). Researchers trained in psychology and other behavioral sciences may be familiar with catalogs of threats to validity that describe prototypical problematic situations (Matthay & Glymour, 2020). These lists can grow rapidly. For external validity alone, Shadish et al. (2002) distinguished five types of threats that include interactions of the causal relationship of interest with specific units, settings, mediators, outcomes, and treatment variations.

Because of these threats, there have been reasonable calls for constraint. Yarkoni (2022), for instance, argued that poor alignment between verbal hypotheses and quantitative inference lies at the heart of many of psychology’s problems; narrow and seemingly arbitrary operationalizations of broad constructs invalidate the intended generalizations. As a remedy against, often implicit, unwarranted generalizations, researchers have proposed to add mandatory “Constraints on generality (COG)” statements to all empirical articles (Simons et al., 2017; Tiokhin et al., 2019). By specifying sample characteristics and assessing its representativeness of wider populations, such COG statements are meant to discipline authors to explicitly state intended generalizations and thereby improve transparency.

These steps toward a more global and generalizable science are overdue. However, under the current framework—with its emphasis on threats to validity, constraints, and the accumulation of cross-cultural samples—only limited progress can be made. Lists of threats are devices that raise awareness of inferential problems, but they are not also solutions. They do not spell out which inferences are warranted and under which assumptions. This leads to the impression that any claim that goes beyond the precise operationalization, population, and historical context of a study overgeneralizes.

From this perspective, it is understandable that researchers are eager to collect rich data sets just to describe what is “out there” (e.g., Barrett, 2020; Rozin, 2001). But even this is not possible without an explicit

framework that licenses generalization. In a cross-cultural context, even “mere description” and simple comparisons rely on usually implicit assumptions that permit moving from sample to population and across populations.

Threats and constraints forbid inference; we require a framework that also licenses inference. Such a framework would inform researchers about the assumptions underlying potential generalizations, assist them in the design of empirical studies, and show them how to construct appropriate statistical procedures. Such a framework already exists and has sparked a “causal revolution” (Pearl, 2018) in computer science and machine learning, but it is not a standard part of training in the behavioral and social sciences. This framework depends on transparent, generative models of research. One key idea is that generalizability does not depend on the presence of sample differences per se or on raw statistical associations. The conditions that license generalization and comparison with other populations depend on the causal relations between variables and the exact mechanisms by which populations differ.

Many cross-cultural scientists already pay close attention to concerns of causal inference and comparison without use of a formal framework (e.g., Norenzayan & Heine, 2005; Pollet et al., 2014). For these researchers, a formalized framework can provide a vocabulary to articulate their concerns and work toward solutions in a more systematic manner.

For instance, many researchers will share the intuition that the demographic breakdown and other relevant factors should be somehow standardized across groups to eliminate potential confounds. A standard approach to dealing with such threats to validity and cross-cultural comparison is to condition on (i.e., adjust or “control” for) any potential confounds such as age, income, or methodological differences by, for example, including such variables as predictors in multiple regression (conditioning on a variable means to analyze the values of other variables for a given, constant value of the conditioned variable). But it is not enough to mechanically control for a set of variables that may vary across populations. One reason is that not all controls are good—adding variables can bias inference as much as it can correct it (Cinelli et al., 2020). An important example is “collider bias,” in which a spurious association between two variables arises when a third variable, which is jointly caused by those variables, is included. As we show below, which variables act as confounds depends on the assumed causal structure and the specific research question. A formal generative framework lets us logically deduce which variables we should—and should not—control for in any cross-cultural comparison. Going beyond the question of which variables to include, it also helps us derive the appropriate statistical estimates that actually align with the scientific goal at hand. Coefficients and parameters themselves are valid measures of difference or causal effect in only the

simplest models (Morgan & Winship, 2015; Rohrer & Arslan, 2021). Knowing a cause means that we can predict the consequences of an intervention (Asteriou & Hall, 2015; Athey & Imbens, 2016; Greene, 2000; Morgan & Winship, 2015; Woodward, 2005), and most causal questions require the construction of “marginal” effects, in which we average the effect of interest over the influence of all other important variables to find out how a dependent variable would change if we intervened on the independent variable. Such “poststratification,” that is, reweighting of model estimates to answer specific causal questions, becomes even more complicated when comparisons are made between societies (Oganisian & Roy, 2021).

In short, there is no universally valid procedure for cross-cultural inference. For each inferential problem, we have to start with a generative causal model that lets us determine the role variables play in the analysis and how to construct statistical summaries that are logically derived from transparent research goals.

In the rest of this article, we outline a formal framework for cross-cultural generalizability based on recent advances in the fields of causal inference and data fusion (Bareinboim & Pearl, 2016; Lundberg et al., 2021; Pearl, 2015; Pearl & Bareinboim, 2014). We apply these established formal tools to commonplace questions in cross-cultural research: (a) description of cultural variation, (b) comparison of causal effects identified through experiments, and (c) measurement equivalence or inequivalence of latent constructs. To help researchers adopt this approach, we provide example causal diagrams and statistical analyses using simulated and real-world cross-cultural data. Finally, we discuss how our framework can assist researchers in planning targeted cross-cultural comparisons and designing more informative studies.

A Causal Framework

A causal framework for cross-cultural research requires us to state (a) what we want to know, that is, the estimand; (b) a generative model of the evidence, that is, a causal model of how the observed data came into existence; (c) a generative model of how populations may differ; and (d) a tailored estimation strategy that allows us to learn from data. We first develop these requirements in general terms. In later sections, we discuss specific examples.

Theoretical and empirical estimands

The starting point for any empirical analysis is the theoretical estimand. This is the target of the analysis derived from theory (for an excellent introduction, see Lundberg et al., 2021). A theoretical estimand consists of a unit-specific quantity and a target population. It is defined outside of any statistical model—not in terms of, for

example, regression coefficients. We may simply be interested in the mean of a variable in a certain population (e.g., probability that individual i chooses the prosocial option in a dictator game, averaged over all individuals i in target population), or we may be interested in the average treatment effect of some independent variable on an outcome in a certain population (e.g., effect of norm prime on probability that individual i chooses prosocial option, averaged over all individuals i in target population; examples inspired by House et al., 2020; see below).

Once the theoretical estimand is set, we need to link it to an empirical estimand. While the theoretical estimand might contain unobservable quantities such as counterfactuals (“What would have been true under different circumstances?”), the empirical estimand is defined solely in terms of observed data. We cannot observe the average probability of prosocial choice for the whole population; however, we can try to estimate it from a sample. We also cannot observe individual-level causal effects, but we may estimate their average by considering observed differences between randomized experimental conditions.

In the context of cross-cultural research, the distinction between theoretical and empirical estimands encourages researchers to explicitly spell out assumptions about how theoretical constructs (e.g., prosociality) can be operationalized in comparable ways across societies (“construct validity”). This issue of measurement equivalence or inequivalence and bias is further discussed in the “Generalizing Latent Constructs: Measurement equivalence or inequivalence” section.

Directed acyclic graphs

A valid link between theoretical and empirical estimand requires causal assumptions. Generative models embody causal assumptions, and there are many forms these models can take. One popular approach is directed acyclic graphs (DAGs). This approach is accessible thanks to its graphical nature, it can be used to develop an intuitive understanding of inferential obstacles, and it can alert researchers to inferential opportunities they had not considered. There are other suitable ways to spell out assumptions (e.g., psychological process models; Farrell & Lewandowsky, 2018), and not all generative models can be formalized with the help of DAGs. But DAGs provide a pragmatic starting point and can be extended to include commonplace issues such as measurement error and missing data (see McElreath, 2020, Chapter 15).

Multiple comprehensive yet accessible introductions to DAGs are available (Elwert, 2013; Pearl et al., 2016; Pearl & Mackenzie, 2018; Rohrer, 2018); thus, we focus only on the essentials. In DAGs, nodes represent variables, and arrows represent causal effects. For example, Figure 1a captures a set of assumptions regarding the associations

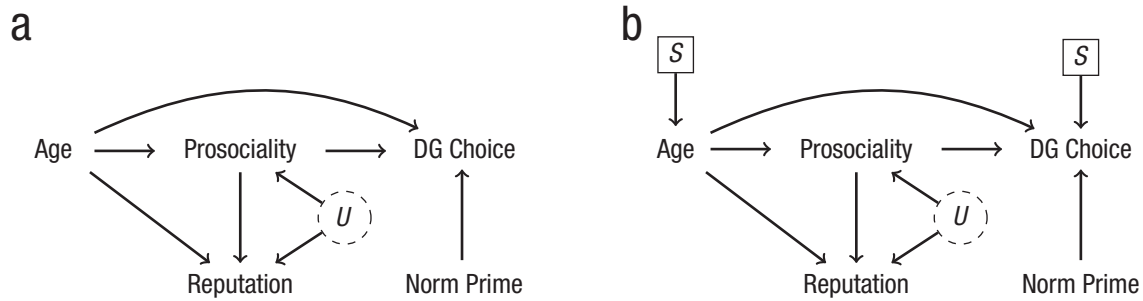


Fig. 1. (a) A simple directed acyclic graph capturing the following assumptions: Age has a direct causal effect on an individual's prosociality, their reputation within their community, and the outcome of the dictator game (DG). Prosociality and reputation share an unobserved common cause, U . Prosociality in turn affects the individual's choice in the dictator game, which is also affected by the randomized norm prime. (b) Selection diagram using selection nodes \boxed{S} to represent the assumption that populations differ both in their age distribution and the effect of norm primes on the choice in the dictator game.

between age, prosociality, reputation, and the outcome of a dictator game. The arrows indicate causal effects that may take any functional form, which includes any possible interaction between variables that jointly affect another variable. Individual paths can be identified by traveling along the arrows connecting any pair of variables. These paths can be broken down into fundamental structures (see Box 1) that determine whether a given path transmits an association between variables and whether the association is causal or noncausal.

Suppose we were interested in the causal influence of prosociality on dictator-game choice in the population from which we randomly drew our sample. If we are willing to assume that the depicted DAG is a causal DAG—which means that it includes all common causes of any pair of variables (Elwert, 2013)—we can algorithmically derive which variables need to be “conditioned” (see Box 1) on to identify the causal effect of interest. In this particular example, the answer is easy. There is only one open noncausal path (see Box 1) between prosociality

Box 1. Elementary Causal Structures

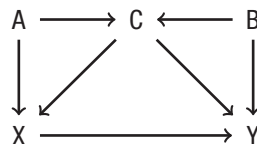
Any path connecting two variables can be broken down into three fundamental causal structures: chains, forks, and inverted forks (Elwert, 2013; Rohrer, 2018).

Chains: $X \rightarrow M \rightarrow Y$. The chain transmits a causal association between X and Y . If we condition on M (the mediator; e.g., through statistical adjustment, sample stratification, or by design), the transmission of the association is blocked.

Forks: $X \leftarrow C \rightarrow Y$. The fork transmits a noncausal association between X and Y . If we condition on C (the confounder), the transmission of the association is blocked.

Inverted forks: $X \rightarrow L \leftarrow Y$. The inverted fork transmits no association. If we condition on L (the collider), a noncausal association between X and Y is transmitted.

A path between X and Y is said to be *d-separated* if it contains a confounder or mediator that has been conditioned on or a collider that has not been conditioned on (Pearl, 1988). This implies that the path will not transmit any association; it is “blocked.” For a statistical procedure to recover a causal association, it must be designed to block any noncausal paths. For example, in the directed acyclic graph below, we wish to measure the causal association between X and Y . There are, however, two noncausal paths that also connect X to Y . The first is $X \leftarrow C \rightarrow Y$. This is a confounder path, and we close it by conditioning on C . The second noncausal path is $X \leftarrow A \rightarrow C \leftarrow B \rightarrow Y$. In this path, the variable C is not a confounder but rather a collider. As a result, this path would normally be closed. But after we condition on C to close the first path, it opens the second path. Therefore, we must also condition on A or B to close this second path. Therefore, a procedure that measures the association between X and Y , stratified by C and B (or A ; but using B also increases precision [Pearl et al. 2016] and may thus be preferable) would measure the causal effect of X on Y .



and dictator-game choice: Prosociality \leftarrow Age \rightarrow Dictator-game choice. Because age is a common cause of prosociality and dictator-game choice, some of the association between both variables is due to this noncausal path. This path can be blocked by conditioning on age (again, see Box 1). Thus, we have discovered a way to link the theoretical estimand (the effect of prosociality on dictator-game choice in our population) to an actual empirical estimand that we can estimate from observable data. If, instead, our theoretical estimand was the effect of the norm prime, no conditioning would be necessary for causal identification: Because the norm prime has been randomized, no backdoor paths can exist (no arrows point into the randomized variable). The simple mean difference between experimental groups would be an empirical estimand that corresponds to the theoretical estimand under the assumptions embodied in the DAG (taking into account other variables that influence the outcome may still be helpful to improve precision).

Our DAG is, of course, incomplete and possibly wrong, in particular when it comes to the nodes that have not been experimentally manipulated. But an incomplete model is still an improvement over no model at all. In the absence of causal assumptions, whether in a DAG or otherwise, no analysis can be scientifically justified. Even an unrealistic DAG can help identify specific problems as well as implicit assumptions underlying more casually drawn causal inferences. Furthermore, such graphs make it easier to contrast the implications of different sets of assumptions that often lie at the heart of scientific disagreements. Throughout this article, we use DAGs in this spirit—as a pragmatic tool to communicate assumptions and improve inference.

Selection diagrams and generalizability

DAGs can be extended to address generalizability through the use of selection diagrams (Pearl & Bareinboim, 2014). When researchers consider multiple populations, selection diagrams allow them to precisely define the local mechanisms by which populations are assumed to differ, as represented by “selection nodes.” Selection nodes are not variables but, rather, indicate which nodes have culture-specific distributions or causal relationships.

Returning to our previous example, in Figure 1b, we added two selection nodes. The $\boxed{S} \rightarrow$ Age node may indicate that populations are characterized by different age distributions, and the other \boxed{S} node may indicate that the populations differ in the weight individuals give to norm primes when making decisions in the dictator game (recall that in a DAG, all variables that jointly affect another variable may interact). The absence of selection nodes in such graphs is of equal importance. It represents the assumption that certain mechanisms are the same across populations. For instance, the diagram in

Figure 1b implies that the development of prosociality with age does not vary among study populations. As shown below, it is this assumed invariance of certain mechanisms that makes generalizations possible.

Once we have a causal selection diagram, we can determine the scope for generalizability using logical rules. We can deduce when and how we can use data from one population to estimate a target quantity in another population, which is the central goal of the literature on transportability and data fusion (Bareinboim & Pearl, 2016; Cinelli & Pearl, 2021; Pearl, 2015; Pearl & Bareinboim, 2014). These logical rules can be compressed in most contexts to a set of simple graphical criteria, allowing us to perform the logic with our eyes (see “Applying the Causal Framework” section).

In cross-cultural settings, the research question often does not directly concern transport. Instead of transporting an estimate from one population to another, we instead have data sampled from multiple populations and want to make sense of the resulting numbers to learn more about whether, how, and why people differ from one another. However, such cross-cultural comparisons are still indirect exercises in transport because to compare distributions or causal effects in different populations, we must calculate what those distributions or effects would be if we changed the population.

Estimation: multilevel regression with poststratification

After establishing the logic of a generalization, one must actually compute it. For explicit generalization from sample to population and comparison across populations, we used multilevel regression with poststratification, a statistical technique that adjusts for differences between a sample population and a target population (Gao et al., 2021; Gelman & Little, 1997; Wang et al., 2015). In a first step, the model uses partial pooling to obtain robust estimates for each “cell” (combination of attributes that we want to condition on; e.g., age/gender groups) taking into account information gained from other cells (Gelman & Hill, 2006; McElreath, 2020). For the data examples below, we used Gaussian processes to obtain estimates for each gender and age group while treating age as a continuous dimension; similar ages were expected to be similar in terms of their prosocial tendencies. In the second step (the poststratification), estimates for all cells are reweighted using the relative frequencies of individuals per cell in the target population (for detailed explanation and model equations, see Appendix A in the Supplemental Material available online; for *Stan* [Carpenter et al., 2017] code used to implement all analyses, see the GitHub repository: <https://github.com/DominikDeffner/Cross-Cultural-Generalizability>).

Multilevel regression with poststratification enables us to learn from data and to project or “generalize” results

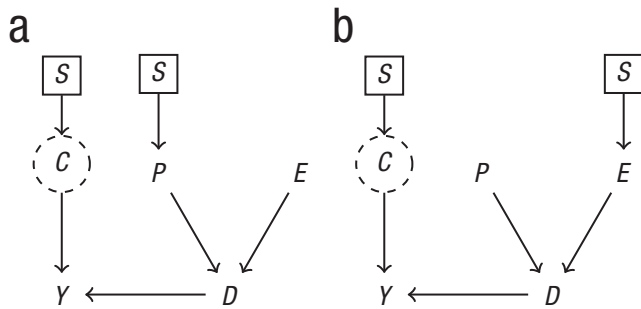


Fig. 2. Different sources of demographic disparities among study samples. Prosociality Y is caused by demography D and unobserved cultural factors C . The sample demography D is caused by population demography P and sampling procedures E . Selection nodes S indicate mechanism by which populations differ. In addition to latent cultural factors, societies can differ in terms of (a) population demography or (b) sampling procedures.

to populations beyond the study sample in a principled way. Which population to use for poststratification depends on the theoretical estimand, the target of inference, and causal assumptions about the data-generating process. Compared with more informal reweighting procedures, multilevel regression with poststratification propagates uncertainty through all steps of analysis and is thus particularly suited for the small samples common in cross-cultural research.

Note that although the use of multilevel regression is not logically required—there are other estimation approaches—the use of poststratification is. The DAGs we describe below mandate poststratification as a logical consequence of their structure. Informal reweighting is only sometimes equivalent to this approach. In every case, the proper way to reweight estimates is a consequence of causal assumptions.

Applying the Causal Framework

To illustrate our approach, we used a large-scale cross-cultural project on societal diversity in prosocial behavior as an empirical case study (House et al., 2020). The researchers administered a binary-choice version of the dictator game as a measure of costly sharing to 255 adults and 833 children from eight populations spanning foragers, small-scale horticulturalists, and urban communities (for demographic composition of samples, see Appendix D, Fig. S3, in the Supplemental Material). Participants were asked to choose between a “self-maximizing” option in which they would keep two rewards or a “prosocial” option in which they would keep one reward and give one to an anonymous peer. Children from six societies were divided into three experimental conditions in which they viewed a short video with normative information before making their choices. These norm primes communicated which behavior was preferable

(“Generous,” “Both OK,” or “Selfish”). We used this rich data set because it exemplifies the state of the art in experimental cross-cultural research and excels with respect to research transparency.

Generalizing description: cross-cultural comparisons and demographic standardization

A basic aim of cross-cultural research is to describe cultural variation. In the simplest case, we might want to compare the prevalence of some institution or behavior across societies. This seemingly innocuous task of “pure description” may actually refer to a number of different research questions that call for different procedures. The example we provide is simplified and focuses on demography, but the point is not about demography. The same logic applies to all comparisons in which populations differ in any known background factors.

Drawing out the causal assumptions. Samples from different sites often differ in terms of their demographic profiles (here, their age and gender distribution), and these demographic variables might in turn affect the distribution of the trait of interest.

How should researchers deal with these differences? The answer depends on the processes that generated the observed disparities. Demographic disparities among samples may result from (a) differences in the actual populations from which the samples are taken or (b) sampling procedures that differ among sites. For example, if we observe that a sample from one site is on average younger than a sample from a second site, this may be because the underlying population is indeed younger. Alternatively, the difference could also result from a comparison of a relatively young convenience sample collected at one site with a full community sample at another site.

These scenarios are depicted in Figure 2. In this figure, an observed outcome Y is influenced by both unobserved cultural factors C and sample composition D . The sample composition is in turn influenced by the true demography P and sampling procedures E (for “experimenter”).

If disparities arise from population differences (Fig. 2a), we can directly compare samples as long as our goal is to simply describe population differences in the focal trait Y regardless of whether they arise from demography or from cultural factors. Adjustment is necessary, however, if we are interested in different comparisons. For example, we may be interested in the counterfactual (i.e., hypothetical) distribution of the trait under comparable demographic profiles: If the two sites had comparable age and gender distributions, would we still observe differences in the trait of interest? This way, researchers

could, for instance, isolate the influence of different cultural factors C while holding constant demographic distributions. Note that such counterfactual comparisons might also correspond to a more substantive theoretical estimand (i.e., the distribution of a trait under a hypothetical intervention that moved individuals to another population; Lundberg et al., 2021).

If disparities among sites arise from different sampling procedures E (Fig. 2b), even the purely descriptive question of observable population differences requires demographic adjustment because sample demographics are systematically biased compared with the population of interest. For example, if the gender of the researcher influenced the gender of voluntary participants, then any differences between societies could be due to a mix of cultural, demographic, and sampling differences. In this case, even large samples do not accurately describe the target populations, and we need to poststratify using information about the population from which samples are taken.

Another scenario, not illustrated in Figure 2, is when a sample is selected on the outcome variable Y itself. For example, if prosocial individuals are more likely to cooperate with the researcher, this is selection on the outcome. In this case, there may be no solution to generalize from sample to population and therefore no way to compare populations. This is perhaps the starkest example of how description depends on causal assumptions.

We turn to real empirical data in the next subsection. However, knowing how to simulate data to validate an analytical strategy is also useful. For a walk-through on a complete simulated data example in which we know the true generative process, see Appendix B in the Supplemental Material. We used multilevel regression with poststratification for the situation in which populations differ in their demographic profile (see Fig. S1, left, in the Supplemental Material) and the complementary situation in which demographic profiles of the populations are identical but genders are sampled unequally because of differences in local sampling procedures (see Fig. S1, right, in the Supplemental Material). In the first case, unadjusted empirical estimates accurately recover true population values, but poststratification can be used for counterfactual comparisons. In the second case, only poststratified estimates accurately recover true population values.

Empirical example. We now turn to our empirical case study on prosociality across societies. Figure 3 shows a comparison between two actual populations included in House et al. (2020), Tanna island in Vanuatu (left) and Berlin in Germany (right). These societies have very different demographic profiles and sample compositions. Here, we were immediately confronted with a pragmatic concern: For many populations, no fine-grained demographic

information is available. Therefore, we had to use the demography of all of Vanuatu instead of only Tanna. This highlights how collecting basic descriptive information about study populations is a crucial first step for any cross-cultural inference.

We divided data into 20 age categories spanning 5 years each and used Gaussian process multilevel regression with poststratification (for gender- and age-specific model estimates, see Appendix D, Fig. S4, in the Supplemental Material). For Tanna, poststratification to either the demographic profile of Vanuatu or of Berlin leaves estimates unchanged (Fig. 3, bottom). This is because there was only a weak effect of age in this sample and the gender distribution was balanced. For the Berlin sample, on the other hand, adjusting for the demographic population profile of Berlin substantially increased the expected amount of prosociality. This is because older individuals in Berlin tended to be more prosocial in their choices and House et al. (2020) focused their data collection on children, which resulted in a much younger sample compared with the underlying population. Drawing the counterfactual comparison for Berlin individuals under the demographic profile of Vanuatu slightly increased the estimate.

How does this compare with the standard approach in which researchers report raw age- and gender-specific estimates for each sample, thereby “controlling” for any differences? The parameter estimates are necessary, but they are not enough. First, the claim that conditioning on age and gender controls for sample differences depends on causal assumptions, as we explained in the previous sections. Second, the distribution of population differences depends not only on the parameters but also on the distributions of age and gender in each target population. A difference in parameters can look large but have little impact on population differences because both the relevant age-gender categories may be too rare to make a large difference and sizable differences on the parameter (e.g., logit) scale may result in minor differences on the outcome (e.g., probability) scale. Only by poststratifying to the outcome scale and to the relevant target population can behavioral differences be compared (Oganisian & Roy, 2021; Rohrer & Arslan, 2021).

Although these examples have been simplified, they highlight the general concern. To accurately describe the prevalence of a trait and compare it across societies, we need to carefully define our theoretical estimand—consisting of unit-specific quantity and target population—and make assumptions about the processes that generate observed disparities in demography or any other potentially significant variable. After a target population is set, refined statistical procedures, such as multilevel regression with poststratification, allow us to generalize observed outcomes to other populations conditional on causal assumptions.

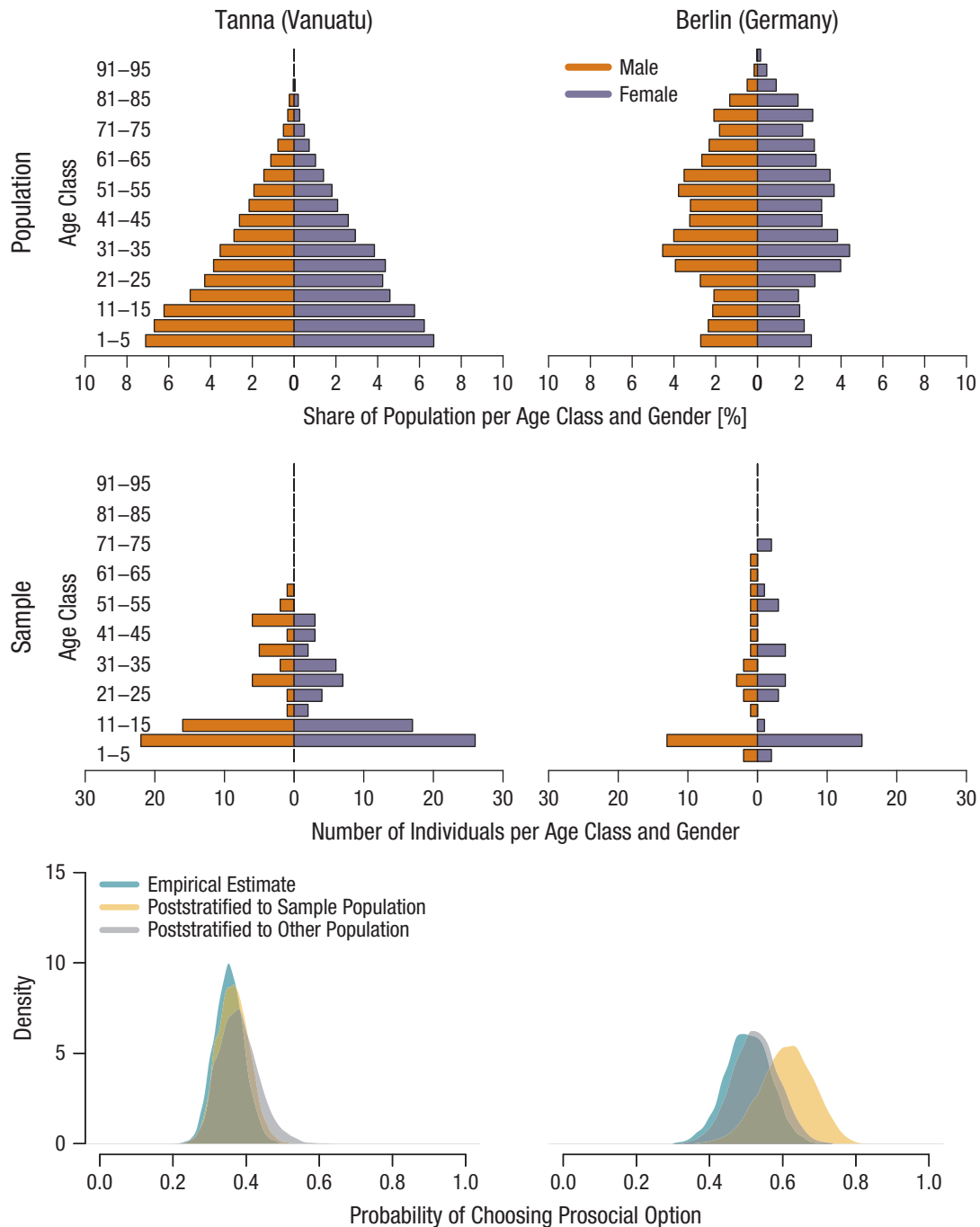


Fig. 3. Data example for demographic standardization comparing prosociality among two populations, (left) Tanna, Vanuatu, and (right) Berlin, Germany. The top row shows demographic profiles of Vanuatu (UN Department of Economic and Social Affairs, World population prospects 2019) and Berlin (Mikrozensus 2020, Amt für Statistik Berlin-Brandenburg); the middle row shows demographic characteristics of study participants from both sites in House et al. (2020); the bottom row shows posterior distributions for probability to choose prosocial option from multilevel regression with poststratification analyses. Blue curves show empirical (unadjusted) estimates, yellow curves are poststratified to be representative of the population from which the sample was drawn, and gray curves are poststratified to demographic profile of other population.

Generalizing experimental results: transportability of causal effects

Many hypotheses in cross-cultural research concern not only the prevalence of a certain trait across societies

but also the causal effect of an independent variable (“exposure,” “treatment”) on a dependent variable (“outcome”). In our example, we were interested in the causal effect of experimental norm primes on prosocial choices in the dictator game (House et al., 2020). Using

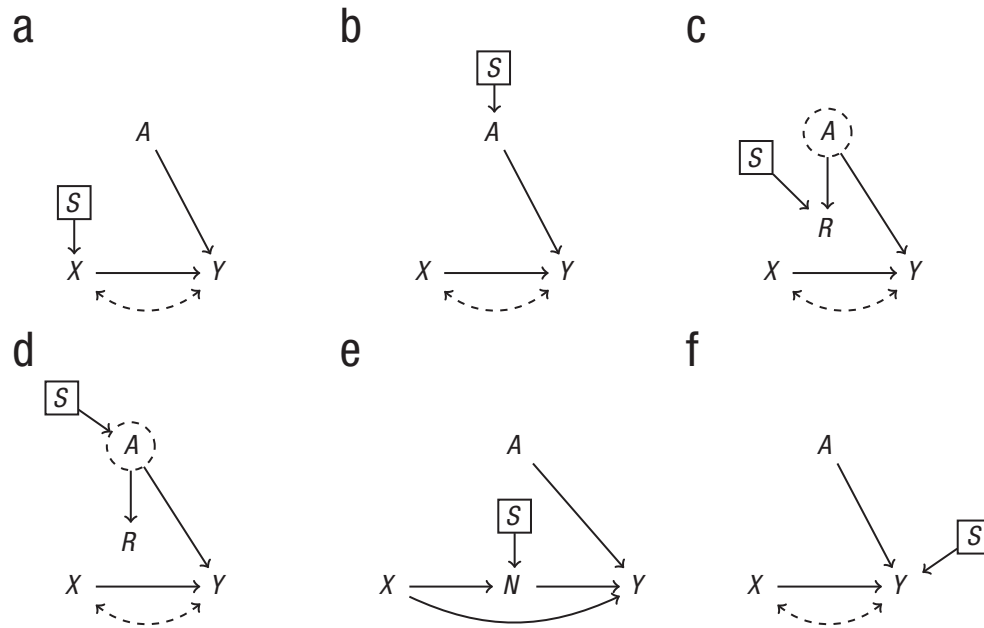


Fig. 4. Scenarios for transportability of causal effects across populations. (a) Normative social information X , which is assumed to differ among populations, causes choice in dictator game Y ; age A modifies effect of X on Y but is invariant across populations; there are also unmeasured confounds between X and Y (indicated by dashed line). (b) Effect-modifier A varies among populations. (c) Age itself is unobserved, but we get to measure reported age R as a proxy. It is assumed that the way people report their ages varies across societies but the underlying age distribution is the same. (d) Ages are reported in the same way across populations, but there are population differences in age distribution. (e) Response in mediator variable, norm activation N , varies across societies. (f) Populations vary in response of outcome variable Y to treatment X . Note that scenarios c, d, and e are described in detail in Appendix C in the Supplemental Material available online.

the transportability framework from causal inference (Pearl, 2015; Pearl & Bareinboim, 2014), we show how causal thinking can be leveraged to generalize and compare causal effects across populations (for formal “S-admissibility” criterion, see Appendix C in the Supplemental Material).

Figure 4 shows selection diagrams for different scenarios varying in terms of scope and procedures for generalizations. They encode different sets of assumptions about the local mechanisms that cause populations to differ. We consider a situation in which normative social information X and age A jointly cause choice in dictator game Y . Note these DAGs represent the “pretreatment” situation, which means X has not yet been experimentally set to a particular value. After X is manipulated through norm-prime videos, all arrows entering X (i.e., all “backdoor” paths) are deleted because the experimentalist is now the sole cause of X . This allows us to estimate the causal effect from observed group differences. An experiment is necessary because we assume unobserved confounds—represented by dashed arrows—that influence both normative social information and prosocial choices (e.g., societies that strongly emphasize prosociality may be structured such that normative information is salient but also encourage prosociality through other means).

Differences in independent/treatment variable. In Figure 4a, populations differ in the distribution of normative social information X . This could mean, for instance, that in some societies, individuals frequently encounter cultural narratives highlighting the importance of prosociality in their everyday life. As we have just shown, treatment randomization used in the experimental study cancels out such differences. As a consequence, the causal effect $X \rightarrow Y$ is directly transportable or generalizable to other populations. In general, all selection nodes pointing into the independent variable (or other arrows that are removed in the X -manipulated graph) can be ignored (Pearl & Bareinboim, 2014).

Differences in effect modifiers. The scenario depicted in Figure 4b is more interesting. Here, we assume population differences in age. Because age modifies (or “moderates”) the effect of normative information on choices (i.e., the effect of norm primes is assumed to be different for different ages) and the age distribution varies across populations, we cannot simply generalize the observed causal effect from one population to another. However, if age is assumed to affect the influence of norm primes in the same way across populations, we can estimate the age-specific effect of X on Y from experimental data and

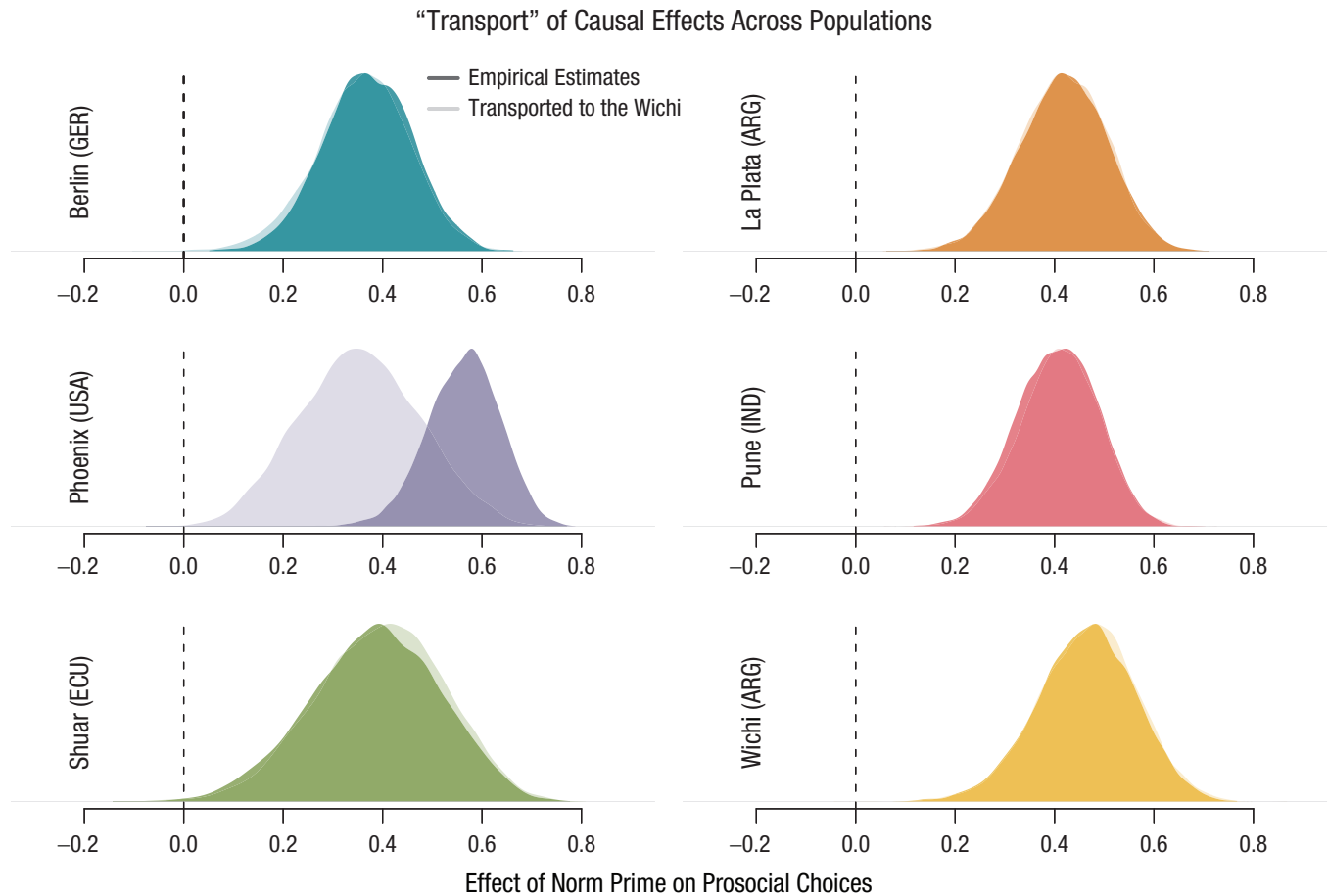


Fig. 5. Data example for transport of causal effects across societies. Empirical estimates (dark colors) and estimates transported to the Wichí in Argentina (transparent colors) for causal effect of norm primes (“Generous” vs. “Selfish”) on prosocial choices in the dictator game in six different societies included in House et al. (2020). Estimates are calculated as the age-specific differences in the probability to choose the prosocial option in both conditions averaged over the age distribution in the target population.

generalize by adjusting for the age distribution of the target population.

The transport approach not only allows principled claims about the generalization of causal effects to new populations, it can also be employed to compare estimates from multiple populations from which experimental data are available. To determine whether observed group differences in causal effects reflect “real” cultural differences (i.e., differences we cannot, yet, explain through other variables) or are due to sampling variation or differences in known effect modifiers, researchers need to make explicit assumptions about the causal processes that generate the data.

Figure 5 shows a data example for the transport of age-specific causal effects across populations (House et al., 2020). Because of experimental manipulation, the causal effect of norm primes X on prosocial choices Y , our estimand, can be estimated from the difference in the probability to choose the prosocial option in both experimental conditions (“Generous” vs. “Selfish”). Dark colors show empirical estimates of this causal effect from

six different societies included in House et al. (2020). Across all societies, posterior densities lay well above 0. This means that individuals who watched the “Generous” prime video were substantially more likely to choose the prosocial option in the dictator game compared with individuals who watched the “Selfish” prime video; the strongest effect was observed in the sample from Phoenix, Arizona, United States.

To adjust for differences in the age distribution as a potential effect modifier, we estimated age-specific causal effects in each society and, as an example, adjusted estimates to the demographic profile of the Wichí in Argentina. Transparent colors in Figure 5 show such counterfactual estimates for the effect of norm primes in each society assuming it had the same demographic composition as the sample from the Wichí. Although estimates remained largely unchanged for most societies, the effect for Phoenix became substantially smaller and more uncertain. This is because in Phoenix, age strongly modifies the effect of norm primes: Younger children were more influenced by norm primes than older children. The Phoenix sample is,

on average, almost 3 years younger than the Wichí sample, so estimates of the causal effect need to be adjusted to apply correctly to the Wichí demographic situation. On the basis of a comparison of naive empirical estimates, researchers might have wrongly concluded that norm primes have a particularly strong effect in Phoenix for some age-invariant cultural reason; transported estimates instead suggest that the larger effect is attributable to (potentially culturally determined) effect modification in combination with the younger sample. Adjustment for potential effect modifiers such as age, therefore, allows researchers to compare causal effects on an equal footing.

To aid understanding, most examples have been relatively straightforward, so some researchers might wonder what they gain from this causal approach compared with more informal ways to standardize and compare estimates across groups. Building up from those fundamental units, in Appendix C in the Supplemental Material, we describe more complicated situations in which implied generalizations and transport formulas could hardly be obtained by intuition alone.

In particular, Appendix C in the Supplemental Material introduces scenarios in which we did not observe the true effect modifier, biological age, but only some proxy, such as reported age R , which is observed to vary across populations (Figs. 4c and 4d). Because different scenarios will generate identical data distributions, the correct procedure will depend solely on causal assumptions. In Appendix C in the Supplemental Material, we further discuss situations in which a mechanism mediating the effect of X on Y differs among societies (Fig. 4e), which requires a more sophisticated—yet algorithmically derivable—generalization formula.

“Impossibility” of generalizations. Finally, if a selection node is pointing directly into outcome variable Y (Fig. 4f), no generalizations are possible because there is no immediate way to account for the source of disparity among populations (for “S-admissibility” criterion, see Appendix C in the Supplemental Material). This would be the case if unobserved population differences directly modify the effect (e.g., Oyserman & Lee, 2008, found that individualism-collectivism primes do not function in comparable ways across societies) or if the form of age modification varies between sites. However, even such “impossible” cases might allow generalizations and comparisons if researchers make additional assumptions, for example, if we have additional knowledge about the mechanisms causing the outcome variable and if only some of these differ among populations (for an example analyzing effects of Vitamin A supplementation on childhood mortality, see Cinelli & Pearl, 2021).

These examples demonstrate that the generalizability of experimental effects does not depend on the presence of population differences per se but on the exact

mechanisms by which populations differ. While some differences—especially those concerning the independent variable—are inconsequential for intended generalizations, differences concerning effect modifiers or mediators require statistical adjustment. Differences in the immediate mechanisms causing the outcome render generalizations difficult or even impossible. Such “real” cross-cultural differences may be the result of society-level factors directly influencing the trait of interest, and they present irreducible obstacles to generalization. Whether “real” cultural differences exist or whether they must eventually be explained away by other mechanisms is a topic beyond the scope of this article.

Generalizing latent constructs: measurement equivalence or inequivalence

In all examples so far, we assumed that researchers can readily observe and measure the variables of interest. However, many (cross-cultural) psychologists are particularly interested in the comparison of latent constructs that are not directly observable. For example, researchers typically do not want to learn about dictator-game choices per se but about the underlying psychological constructs (e.g., “prosociality”) that are assumed to generate the observed choices (for potential impacts of cultural context on economic game choices, see e.g., Bond et al., 1982; Lesorogol, 2007; Leung & Bond, 1984; Pisor et al., 2020). In this section, we briefly demonstrate how causal selection diagrams can be used to represent common issues of measurement equivalence or inequivalence in cross-cultural studies; note that this is just a sketch; doing justice to this issue would require a whole article.

Methodologists have long discussed whether and how data generated in cross-cultural research can be interpreted in terms of the presumed underlying processes and constructs. The “equivalence and bias” framework, for instance, differentiates between construct equivalence, metric equivalence, and scalar equivalence (e.g., Van de Vijver & Leung, 2021; Van de Vijver & Tanzer, 2004). Direct comparisons of measurements across societies are justified only if the underlying construct, measurement units, and scale origin are equivalent across societies (i.e., full scalar equivalence).

But we can also approach the problem from a generative perspective. The measurement process can naturally be represented as a causal model of observed item or test scores (Bandalos, 2018; Borsboom et al., 2004). Note that there are alternative models in which constructs are not seen as common causes of manifest variables but as network structures (Borsboom et al., 2021) or organizing principles (Sijtsma, 2006) that connect such variables; however, the implications of such models for generalizability are beyond the scope of this article.

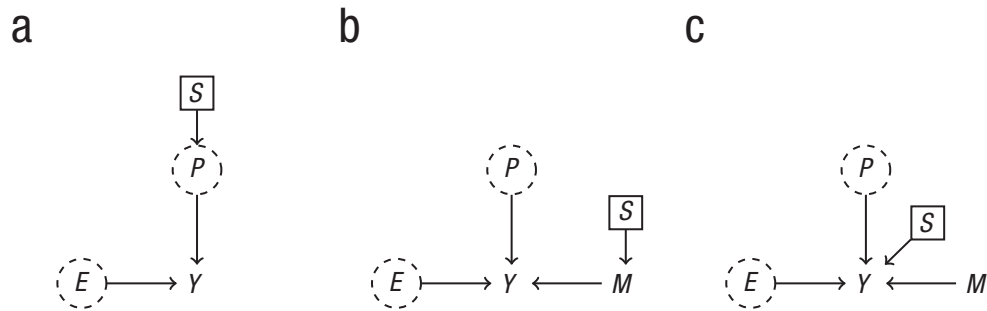


Fig. 6. Causal representation of measurement equivalence or inequivalence across societies. (a) Choice in dictator game Y is caused by latent psychological factor Prosociality P , which varies across populations, and unobserved sources of random error E . (b) Choice in dictator game Y is also influenced by (population-specific) degree of market integration M . (c) The influence of prosociality P , error E , or market integration M on observed choices differs among societies.

For Figure 6a, we assume that an individual’s choice in the dictator game Y is caused by a latent psychological factor P (for “Prosociality”) and unobserved sources of random “error” E . Measurement equivalence in this framework then requires that (a) only the distribution of the latent factor P might vary across communities, (b) P influences Y in the same way everywhere, and (c) there are also no population differences in the unobserved error sources E . These conditions are fulfilled in Figure 6a, so in this case, we would be justified to compare game choices as indicators of latent “prosociality” across communities.

In Figure 6b, choices in the dictator game do not only reflect prosociality and random error but also the degree of market integration M . People who engage more in market activities might be more likely to give a reward to an anonymous peer simply because they are more used to interacting and trading with unknown others, not because they are more prosocial. In this case, choices in the dictator game are not equivalent measures of the latent factor in different societies because they also include the influence of market integration that varies across societies. Nonetheless, following the logic on generalizing description (see “Generalizing Description: Cross-Cultural Comparisons and Demographic Standardization” section), if we have data on market integration for each society, we can use poststratification to adjust for different levels of this variable and arrive at valid comparisons of prosociality and its causes across societies.

Finally, Figure 6c shows a scenario in which the selection node directly points into dictator-game choice Y . This comprises situations of construct inequivalence in which the latent construct itself is not comparable with respect to its influence on manifest behavior but also cases in which the influence of market integration or of unobserved error sources differs among societies. Mirroring the impossibility of transport with selection nodes pointing directly into outcome Y (see Fig. 4f), in any

such case, generalizations and comparisons about latent factors are unwarranted (unless additional assumptions are made). Because there is no way to statistically account for different sources of variation of observed choices Y , we cannot identify the unique influence of the latent state P in equivalent ways across communities.

Using the Causal Framework for Principled Study Design

A causal framework is not only useful for analysis but also aids research design. To connect research designs to selection diagrams, we considered three stereotyped cases: a “maximally diverse” sampling strategy, a “proxy control” approach using phylogenetic distance or shared history, and a “regional comparative” approach that explicitly designs for local causal identification of the mechanisms by which populations differ. We explain each in turn.

A common approach in cross-cultural study design is to aim for maximally diverse populations. If effects can reliably be found across diverse societies, the reasoning goes, researchers are justified in assuming cross-cultural invariance or even universality; differences among samples are interpreted as evidence for either the influence of observed or unobserved cultural factors or methodological differences. By comparing geographically and culturally distant societies, this approach addresses “Galton’s problem,” which describes the pitfalls of drawing inferences from cross-cultural data that are autocorrelated because of shared cultural and historical roots (Naroll, 1965). This rationale guided the construction of the widely used “Standard Cross-Cultural Sample” (Murdock & White, 1969). Figure 7a encodes a scenario in which researchers lack substantive theory on the factors causing a trait Y that varies cross-culturally. Thus, only a selection node is pointing into Y . Because there

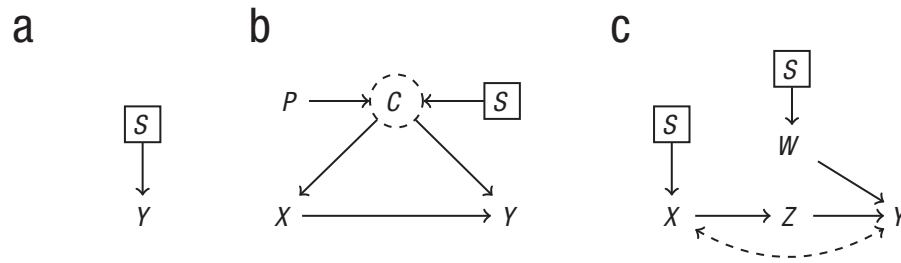


Fig. 7. Different causal scenarios for study design. (a) Unobserved factors cause cross-cultural variation in outcome variable Y . (b) X is a cause of Y , and unobserved cultural variables C that differ between populations influence both X and Y ; phylogenetic relationships P influence C . (c) Causal effect of X is mediated by Z , and W modifies the effect of Z on Y . Dashed arrow represents unobserved common causes.

is no way to separate sources of population differences from the trait itself, there are no theoretical grounds to predict how the trait might vary across populations. In such exploratory scenarios, it is advisable to sample many culturally distinct societies to approach a representative sample of the full range of variation (for description of cultural variation, see “Generalizing Description: Cross-Cultural Comparisons and Demographic Standardization” section). In general, when there is a selection node pointing directly into the outcome variable, researchers must incorporate relatively diverse populations because there are relevant but unknown variables causing population differences. However, potential dimensions of variation across settings, individuals, and societies are effectively infinite. They can never be sampled exhaustively, which reflects the classic problem of induction (Hume, 1739/2003; Sloman & Lagnado, 2005). In addition, although this approach reduces the chance that cross-cultural similarity is due to recent shared influences, it is not a general solution to causal inference because any similarity between distant societies could still be due to unobserved variables.

A generalizable understanding of a given phenomenon, therefore, cannot be based only on the accumulation of data but requires the theory-driven testing of causal assumptions. How even the most rudimentary causal theory helps increase generalizability can be seen in Figure 7b, in which researchers have identified an explanatory variable X . If researchers can find an identification strategy to estimate the causal effect $X \rightarrow Y$, they can leverage this causal knowledge to enhance generalizability following the transport approach outlined in “Generalizing Experimental Results: Transportability of Causal Effects” section. The problem is that unobserved cultural variables C , which differ between populations, influence both X and Y and thus confound the causal effect. One approach is to try to model the covariation among populations that arises from such unobserved confounds. Variables such as geographic, linguistic, or cultural distance P can be used as proxies

to control for unmeasured common causes of similarity. The notion is that populations closer in space or cultural history share more unmeasured common causes. This can permit causal investigation of, for example, ecological and demographic factors in otherwise opportunistic collections of societies. Various cultural and linguistic phylogenetic methods try to implement this strategy (for detailed examples, see McElreath, 2020, Section 14.5). This approach makes strong causal assumptions about the nature of confounding and our ability to measure shared history. However, strong assumptions are always necessary in observational settings. What is important is that the assumptions are transparent and logically connected to data analysis.

Finally, Figure 7c shows a scenario in which researchers have developed more mechanistic theory including additional variables lying on the causal paths between selection nodes and outcome; this provides more principled expectations about the mechanisms generating population disparities. Specifically, there is an intermediate variable Z mediating the effect of X on Y and another variable W that modifies the effect of Z on Y . If this DAG is assumed, there is no selection node pointing into Y anymore, and thus researchers can explain all population differences in the focal trait on the basis of the joint causal effects of other variables. A research design that attempts to address causation directly is the “regional comparative” approach. In this approach, researchers explicitly target closely related societies that differ only in key variables of interest (Boas, 1896; Johnson, 1991). By holding other factors constant, such “quasi-experimental” comparisons among regional populations or subpopulations allow researchers to isolate the effect of a variable of interest and facilitate causal inference. This strategy is similar to difference-in-difference (Lechner, 2011) and regression-discontinuity designs (Imbens & Lemieux, 2008; Lee & Lemieux, 2010). A classic example of the approach is the Culture and Ecology in East Africa Project that compared samples from four different ethnic groups, each of which comprised neighboring pastoralist and horticulturalist

communities in different but adjacent ecologies (Edgerton, 1971; Goldschmidt, 1965). Although differences among ethnic groups are hard to interpret, differences between neighboring communities in each ethnic group are arguably due to local ecological and economic differences (for more recent examples, see Glowacki & Molleman, 2017; Mattison et al., 2016).

To summarize, because of the problem of induction, generalizability can never be determined through the accumulation of cross-cultural data alone and requires the development of formal theory to accompany and guide cross-cultural data collection (Muthukrishna & Henrich, 2019). The maximally informative research design depends on the state of mechanistic understanding of the phenomenon of interest. By explicitly stating and refining the causal assumptions underlying population differences, researchers can target maximally informative cross-cultural comparisons and generate results that are not only grounded in theory but also generalizable beyond the immediate study samples.

Conclusions

More diverse samples are urgently needed, but they bring forth new conceptual challenges for description, generalization, and comparison. The accumulation of large cross-cultural data sets in combination with lists of threats to validity allows only limited progress. What is needed in addition is a structural-causal-modeling framework. An explicit causal framework empowers researchers by providing a way to plan cross-cultural comparisons, implement and justify analyses, and determine which interpretations are warranted under which sets of assumptions. It also provides a powerful way to critically and fairly evaluate the studies of others and to formally represent sources of disagreement. An effective critique should aim for the same causal clarity as an effective study. When an original study lacks causal clarity, an effective critique may identify which causal model is implied by the analysis and subsequently assess the plausibility of specific elements.

Researchers in various fields already apply methods that address some of the concerns we discussed above. For example, political scientists and sociologists apply demographic standardization (e.g., Kitagawa decomposition) to estimate effects of interventions for counterfactual populations (Acharya et al., 2016; Ciocca Eller & DiPrete, 2018; Kitagawa, 1955; Mize, 2016; Preston et al., 2000; Ross et al., 2021; Storer et al., 2020). Anthropologists calculate age-corrected values to standardize across populations (Borgerhoff Mulder et al., 2009; Jaeggi et al., 2021; Mattison et al., 2016; and Rowan et al., 2021). Economists calculate average treatment effects and marginal effects that can take into account effect modification by demographic variables (Asteriou & Hall, 2015; Athey

& Imbens, 2016; Greene, 2000; Morgan & Winship, 2015), and the Heckman correction is applied to account for nonrandom sample selection (Heckman, 1976, 1979; Puhani, 2000). And even simple regression controls can account for population differences in background factors in some limited situations.

The framework we champion—poststratification and transport based on causal graphs—goes beyond these partial solutions. It is explicit about the target of inference and the assumptions that justify the analysis; it logically derives statistical procedures from a generative causal model. Therefore, it is more general and unifies a large number of inferential concerns (e.g., confounding, selection bias, standardization, generalization) in a common framework. Likewise, the estimation strategy that we propose—multilevel regression with poststratification—is very flexible. It allows to project estimates to arbitrary target populations and can account for any number of variables and functional relationships between them. In contrast, simply including age and gender as covariates in multiple regression assumes that all relationships are linear and estimates population differences holding covariates constant at an arbitrary level. Under the right circumstances, this standard approach might tell us something about differences between observed samples but does not enable us to generalize findings to the sample populations (in case of sampling differences) and other populations in a reasoned way.

It is quite obvious that all the scenarios we presented were oversimplified. An explicit causal-inference framework makes it (at times painfully) transparent how strong the assumptions are that we need to arrive at substantive conclusions and how little we collectively know about many real-world phenomena. But this is no reason to embrace the status quo that often avoids causal language (Grosz et al., 2020)—assumptions do not disappear just because we ignore them. Cross-cultural research is daunting, and strong conclusions require strong methods for data collection, its description, and its analysis. A structural causal framework encourages researchers to explicitly spell out their assumptions, removing verbal ambiguity and facilitating communication, and it calls for a cumulative approach to science as one study's findings become the scaffolding assumptions of the next.

Transparency

Action Editor: Mijke Rhemtulla

Editor: Daniel J. Simons

Author Contributions

D. Deffner and R. McElreath conceived the project. D. Deffner wrote the simulations and performed the analyses for the data examples. D. Deffner, J. M. Rohrer, and R. McElreath wrote the manuscript. Conceptualization: D. Deffner, J. M. Rohrer, R. McElreath. Data curation: D. Deffner. Formal

analysis: D. Deffner. Investigation: D. Deffner, J. M. Rohrer, R. McElreath. Methodology: D. Deffner, J. M. Rohrer, R. McElreath. Software: D. Deffner. Supervision: R. McElreath. Visualization: D. Deffner. Writing-original draft: D. Deffner, J. M. Rohrer, and R. McElreath. All of the authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work has been funded by the Max Planck Society.

Open Practices

Open Data: <https://github.com/DominikDeffner/Cross-Cultural-Generalizability>


Open Materials: <https://github.com/DominikDeffner/Cross-Cultural-Generalizability>

Preregistration: not applicable

All data and materials have been made publicly available via GitHub and can be accessed at <https://github.com/DominikDeffner/Cross-Cultural-Generalizability>. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>.



ORCID iD

Dominik Deffner  <https://orcid.org/0000-0002-1649-3861>

Acknowledgments

We thank members of the Department for Human Behavior, Ecology, and Culture and the Department of Comparative Cultural Psychology at the Max Planck Institute for Evolutionary Anthropology in Leipzig for constructive discussions and criticisms that helped improve this article.

Supplemental Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/suppl/10.1177/25152459221106366>

References

- Acharya, A., Blackwell, M., & Sen, M. (2016). Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review*, *110*(3), 512–529.
- Apicella, C., Norenzayan, A., & Henrich, J. (2020). Beyond weird: A review of the last decade and a look ahead to the global laboratory of the future. *Evolution and Human Behavior*, *41*(5), 319–329.
- Asteriou, D., & Hall, S. G. (2015). *Applied econometrics*. Macmillan International Higher Education.
- Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64.
- Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J.-F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences, USA*, *117*(5), 2332–2337.
- Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Publications.
- Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences, USA*, *113*(27), 7345–7352.
- Barrett, H. C. (2020). Deciding what to observe: Thoughts for a post-weird generation. *Evolution and Human Behavior*, *41*(5), 445–453.
- Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist*, *37*(3), 245–257. <https://doi.org/10.1037/0003-066X.37.3.245>
- Boas, F. (1896). The limitations of the comparative method of anthropology. *Science*, *4*(103), 901–908.
- Bond, M. H., Leung, K., & Wan, K. C. (1982). How does cultural collectivism operate? The impact of task and maintenance contributions on reward distribution. *Journal of Cross-Cultural Psychology*, *13*(2), 186–200.
- Borgerhoff Mulder, M., Bowles, S., Hertz, T., Bell, A., Beise, J., Clark, G., Fazzio, I., Gurven, M., Hill, K., Hooper, P. L., Irons, W., Kaplan, H., Leonetti, D., Low, B., Marlowe, F., McElreath, R., Naidu, S., Nolin, D., Piraino, P., . . . Weissner, P. (2009). Intergenerational wealth transmission and the dynamics of inequality in small-scale societies. *Science*, *326*(5953), 682–688. <https://doi.org/10.1126/science.1178336>
- Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., Costantini, G., Isvoranu, A.-M., Wysocki, A. C., van Borkulo, C. D., van Bork, R., & Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, *1*, Article 58. <https://doi.org/10.1038/s43586-021-00055-w>
- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Broesch, T., Crittenden, A. N., Beheim, B. A., Blackwell, A. D., Bunce, J. A., Collieran, H., Hagel, K., Kline, M., McElreath, R., Nelson, R. G., Pisor, A. C., Prall, S., Pretelli, I., Purzycki, B., Quinn, E. A., Ross, C., Scelza, B., Starkweather, K., Stieglitz, J., & Mulder, M. B. (2020). Navigating cross-cultural research: Methodological and ethical considerations. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1935), Article 20201245. <https://doi.org/10.1098/rspb.2020.1245>
- Calder, B. J., Phillips, L. W., & Tybout, A. M. (1983). Beyond external validity. *Journal of Consumer Research*, *10*(1), 112–114.
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312. <https://doi.org/10.1037/h0040950>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A.

- (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1–32. <https://doi.org/10.18637/jss.v076.i01>
- Cinelli, C., Forney, A., & Pearl, J. (2020). *A crash course in good and bad controls*. SSRN. <http://dx.doi.org/10.2139/ssrn.3689437>
- Cinelli, C., & Pearl, J. (2021). Generalizing experimental results by leveraging knowledge of mechanisms. *European Journal of Epidemiology*, 36, 149–164. <https://doi.org/10.1007/s10654-020-00687-4> 1–16.
- Ciocca Eller, C., & DiPrete, T. A. (2018). The paradox of persistence: Explaining the Black-White gap in bachelor's degree completion. *American Sociological Review*, 83(6), 1171–1214.
- Clancy, K. B., & Davis, J. L. (2019). Soylent is people, and WEIRD is white: Biological anthropology, whiteness, and the limits of the WEIRD. *Annual Review of Anthropology*, 48, 169–186. <https://doi.org/10.1146/annurev-anthro-102218-011133>
- Cock, T., & Campbell, D. (1976). The design and conduct of quasi-experiments and true experiments in field setting. In M. D. Dunnette (Eds.), *Handbook of industrial and organizational psychology* (pp. 223–326). Rand McNally.
- Curtin, C. M., Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M., Fitzpatrick, S., Gurven, M., Kanovsky, M., Kushnick, G., Laurence, S., Pisor, A., Scelza, B., Stich, S., Rueden, C., & Henrich, J. (2020). Kinship intensity and the use of mental states in moral judgment across societies. *Evolution and Human Behavior*, 41(5), 415–429. <https://doi.org/10.1016/j.evolhumbehav.2020.07.002>
- Edgerton, R. B. (1971). *The individual in cultural adaptation: A study of four East African peoples*. University of California Press.
- Elwert, F. (2013). Graphical causal models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 245–273). Springer.
- Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4), 1645–1692.
- Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.
- Gächter, S., & Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, 531(7595), 496–499.
- Gao, Y., Kennedy, L., Simpson, D., & Gelman, A. (2021). Improving multilevel regression and poststratification with structured priors. *Bayesian Analysis*, 16(3), 719–744. <https://doi.org/10.1214/20-BA1223>
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
- Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23(2), 127–135.
- Ghai, S. (2021). It's time to reimagine sample diversity and retire the WEIRD dichotomy. *Nature Human Behaviour*, 5(8), 971–972. <https://doi.org/10.1038/s41562-021-01175-9>
- Glowacki, L., & Molleman, L. (2017). Subsistence styles shape human social learning strategies. *Nature Human Behaviour*, 1, Article 0098. <https://doi.org/10.1038/s41562-017-0098>
- Goldschmidt, W. (1965). Theory and strategy in the study of cultural adaptability. *American Anthropologist*, 67(2), 402–408.
- Greene, W. H. (2000). *Econometric analysis* (4th ed., International ed.). Prentice Hall.
- Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, 15(5), 1243–1255.
- Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In S. V. Berg (Ed.), *Annals of economic and social measurement* (Vol. 5, no. 4, pp. 475–492). NBER. <http://www.nber.org/chapters/c10491>
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, 47(1), 153–161. <https://doi.org/10.2307/1912352>
- Henrich, J. (2020). *The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous*. Farrar, Straus and Giroux.
- Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review*, 91(2), 73–78.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain Sciences*, 33(2–3), 111–135. <https://doi.org/10.1017/S0140525X10000725>
- House, B. R., Kanngiesser, P., Barrett, H. C., Broesch, T., Cebicoglu, S., Crittenden, A. N., Erut, A., Lew-Levy, S., Sebastian-Enesco, C., Smith, A. M., Yilmaz, S., & Silk, J. B. (2020). Universal norm psychology leads to societal diversity in prosocial behaviour and development. *Nature Human Behaviour*, 4(1), 36–44. <https://doi.org/10.1038/s41562-019-0734-z>
- Hume, D. (2003). *A treatise of human nature*. Dover Publications. (Original work published 1739).
- Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2), 615–635.
- Jaeggi, A. V., Blackwell, A. D., von Rueden, C., Trumble, B. C., Stieglitz, J., Garcia, A. R., Kraft, T. S., Beheim, B. A., Hooper, P. L., Kaplan, H., & Gurven, M. (2021). Do wealth and inequality associate with health in a small-scale subsistence society? *Elife*, 10, Article e59437. <https://doi.org/10.7554/eLife.59437>
- Johnson, A. (1991). Regional comparative field research. *Behavior Science Research*, 25(1–4), 3–22.
- Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., Ndukaihe, I. L., Bloxson, N. G., Lewis, S. C., Foroni, F., Willis, M. L., Cubillas, C. P., Vadiillo, M. A., Turiegano, E., Gilead, M., Simchon, A., Saribay, S. A., Owsley, N. C., Jang, C., . . . Coles, N. A. (2021). To which world regions does the valence–dominance model of social perception apply? *Nature Human Behaviour*, 5(1), 159–169. <https://doi.org/10.1038/s41562-020-01007-21>
- Kitagawa, E. M. (1955). Components of a difference between two rates. *Journal of the American Statistical Association*, 50(272), 1168–1194.

- Lechner, M. (2011). *The estimation of causal effects by difference-in-difference methods*. Now.
- Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2), 281–355.
- Lesorogol, C. K. (2007). Bringing norms in: The role of context in experimental dictator games. *Current Anthropology*, 48(6), 920–926.
- Leung, K., & Bond, M. H. (1984). The impact of cultural collectivism on reward allocation. *Journal of Personality and Social Psychology*, 47(4), 793–804. <https://doi.org/10.1037/0022-3514.47.4.793>
- Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What is your estimand? Defining the target quantity connects statistical evidence to theory. *American Sociological Review*, 86(3), 532–565.
- Masuda, T., & Nisbett, R. E. (2001). Attending holistically versus analytically: Comparing the context sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology*, 81(5), 922–934. <https://doi.org/10.1037/0022-3514.81.5.922>
- Matthay, E. C., & Glymour, M. M. (2020). A graphical catalog of threats to validity: Linking social science with epidemiology. *Epidemiology*, 31(3), 376–384. <https://doi.org/10.1097/EDE.0000000000001161>
- Mattison, S. M., Beheim, B., Chak, B., & Buston, P. (2016). Offspring sex preferences among patrilineal and matrilineal Mosuo in Southwest China revealed by differences in parity progression. *Royal Society Open Science*, 3(9), Article 160526. <https://doi.org/10.1098/rsos.160526>
- McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC Press.
- Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., Hopkins, E. J., Howard, R. M., Hartshorne, J. K., Jennings, M. V., Simson, J., Bainbridge, C. M., Pinker, S., O'Donnell, T. J., Krasnow, M. M., & Glowacki, L. (2019). Universality and diversity in human song. *Science*, 366(6468), Article eaax0868. <https://doi.org/10.1126/science.aax0868>
- Mize, T. D. (2016). Sexual orientation in the labor market. *American Sociological Review*, 81(6), 1132–1160.
- Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.
- Murdock, G. P., & White, D. R. (1969). Standard cross-cultural sample. *Ethnology*, 8(4), 329–369.
- Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, educated, industrial, rich, and democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science*, 31(6), 678–701. <https://doi.org/10.1177/0956797620916782>
- Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, 3(3), 221–229.
- Naroll, R. (1965). Galton's problem: The logic of cross-cultural analysis. *Social Research*, 32(4), 428–451.
- Nisbett, R. E., & Miyamoto, Y. (2005). The influence of culture: Holistic versus analytic perception. *Trends in Cognitive Sciences*, 9(10), 467–473.
- Norenzayan, A., & Heine, S. J. (2005). Psychological universals: What are they and how can we know? *Psychological Bulletin*, 131(5), 763–784. <https://doi.org/10.1037/0033-2909.131.5.763>
- Oganisian, A., & Roy, J. A. (2021). A practical introduction to Bayesian estimation of causal effects: Parametric and nonparametric approaches. *Statistics in Medicine*, 40(2), 518–551.
- Oyserman, D., & Lee, S. W. (2008). Does culture influence what and how we think? Effects of priming individualism and collectivism. *Psychological Bulletin*, 134(2), 311–342.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.
- Pearl, J. (2015). Generalizing experimental findings. *Journal of Causal Inference*, 3(2), 259–266.
- Pearl, J. (2018). *Theoretical impediments to machine learning with seven sparks from the causal revolution*. arXiv. <https://doi.org/10.48550/arXiv.1801.04016>
- Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4), 579–595. <https://doi.org/10.1214/14-STS486>
- Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.
- Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.
- Pisor, A. C., Gervais, M. M., Purzycki, B. G., & Ross, C. T. (2020). Preferences and constraints: The value of economic games for studying human behaviour. *Royal Society Open Science*, 7(6), Article 192090. <https://doi.org/10.1098/rsos.192090>
- Pollet, T. V., Tybur, J. M., Frankenhuis, W. E., & Rickard, I. J. (2014). What can cross-cultural correlations teach us about human nature? *Human Nature*, 25(3), 410–429.
- Preston, S., Heuveline, P., & Guillot, M. (2000). *Demography: Measuring and modeling population processes*. Wiley-Blackwell.
- Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, 14(1), 53–68.
- Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1), 27–42.
- Rohrer, J. M., & Arslan, R. C. (2021). Precise answers to vague questions: Issues with interactions. *Advances in Methods and Practices in Psychological Science*, 4(2). <https://doi.org/10.1177/25152459211007368>
- Ross, C. T., Winterhalder, B., & McElreath, R. (2021). Racial disparities in police use of deadly force against unarmed individuals persist after appropriately benchmarking shooting data on violent crime rates. *Social Psychological and Personality Science*, 12(3), 323–332.
- Rowan, C. J., Eskander, M. A., Seabright, E., Rodriguez, D. E., Linares, E. C., Gutierrez, R. Q., Adrian, J. C., Cummings, D., Beheim, B., Tolstrup, K., Achrekar, A., Kraft, T., Michalik, D. E., Miyamoto, M. I., Allam, A. H., Wann, L. S., Narula, J., Trumble, B. C., Stieglitz, J., . . . Gurven, M. D. (2021). Very low prevalence and incidence of atrial fibrillation among Bolivian forager-farmers. *Annals of Global Health*, 87(1), Article 18. <https://doi.org/10.5334/aogh.3252>
- Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, 5(1), 2–14.

- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*/William R. Shadish, Thomas D. Cook, Donald T. Campbell. Houghton Mifflin.
- Sijtsma, K. (2006). Psychometrics in psychological research: Role model or partner in science? *Psychometrika*, *71*(3), 451–455.
- Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. <https://doi.org/10.1177/1745691617708630>
- Sloman, S. A., & Lagnado, D. (2005). The problem of induction. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 95–116). Cengage Learning.
- Smaldino, P. E., Lukaszewski, A., von Rueden, C., & Gurven, M. (2019). Niche diversity can explain cross-cultural differences in personality structure. *Nature Human Behaviour*, *3*(12), 1276–1283.
- Storer, A., Schneider, D., & Harknett, K. (2020). What explains racial/ethnic inequality in job quality in the service sector? *American Sociological Review*, *85*(4), 537–572.
- Tiokhin, L., Hackman, J., Munira, S., Jesmin, K., & Hruschka, D. (2019). Generalizability is not optional: Insights from a cross-cultural study of social discounting. *Royal Society Open Science*, *6*(2), Article 181386. <https://doi.org/10.1098/rsos.181386>
- Urassa, M., Lawson, D. W., Wamoyi, J., Gurm, E., Gibson, M. A., Madhivanan, P., & Placek, C. (2021). Cross-cultural research must prioritize equitable collaboration. *Nature Human Behaviour*, *5*, 668–671 (2021). <https://doi.org/10.1038/s41562-021-01076-x>
- Van de Vijver, F., & Leung, K. (2021). *Methods and data analysis for cross-cultural research* (Vol. 116). Cambridge University Press.
- Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *54*(2), 119–135.
- Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, *31*(3), 980–991.
- Winer, R. S. (1999). Experimentation in the 21st century: The importance of external validity. *Journal of the Academy of Marketing Science*, *27*(3), Article 349. <https://doi.org/10.1177/0092070399273005>
- Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford University Press.
- Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, Article E1. <https://doi.org/10.1017/S0140525X20001685>