Köln Z Soziol
https://doi.org/10.1007/s11577-025-01045-6

Kölner Zeitschrift für Soziologie und Sozialpsychologie

RESEARCH ARTICLE

Bringing External Validity into Sociological Research

Richard Breen · **Guanghui Pan**

Received: 25 February 2025 / Accepted: 5 December 2025

Abstract The so-called causal revolution that has spread through economics and into adjacent social sciences, including sociology, has been very much concerned with developing methods by which to arrive at credible causal estimates, especially in nonexperimental, observational settings. In the language of experiments, it has focussed on internal validity. But much less attention has been paid to external validity, that is, whether a causal relationship holds in situations other than the one in which it was found. Thinking about external validity obliges us to consider why we are trying to estimate causal relationships in the first place, and what we think they are for. In this paper we discuss in greater detail what we mean by internal and external validity and set out the assumptions required for causal estimates to have both. We consider some examples from sociological research and the challenges to external validity that they illustrate. We urge sociologists, especially those engaged in nonexperimental research, to pay more attention to external validity. But we also stress that this is important not only for causal research but also for other research, and we illustrate issues of external validity that may arise in a wide variety of noncausal studies. We conclude with some practical suggestions and remarks concerning some of the general issues that arise from our work.

Keywords Relevance · Generalisability · Qualitative and quantitive sociological reseach · Causality · Internal validity

✉ R. Breen
Nuffield College
Oxford, UK
E-Mail: richard.breen@nuffield.ox.ac.uk

G. Pan
Department of Sociology, University of Oxford
Oxford, UK
E-Mail: guanghui.pan@sociology.ox.ac.uk

Published online: 12 January 2026

Externe Validität in der soziologischen Forschung

Zusammenfassung Die sogenannte „kausale Revolution“, die sich in den Wirtschaftswissenschaften und anderen Sozialwissenschaften, darunter auch der Soziologie, verbreitet hat, befasst sich vor allem mit der Entwicklung von Methoden, mit denen sich glaubwürdige kausale Schätzungen erzielen lassen, insbesondere in nichtexperimentellen Settings. In der Sprache der experimentellen Forschung liegt der Schwerpunkt dabei auf der internen Validität. Weitaus weniger Aufmerksamkeit wurde der externen Validität geschenkt, also der Frage, ob ein kausaler Zusammenhang auch in Kontexten außerhalb desjenigen besteht, in dem er ermittelt wurde. Wenn wir uns mit der externen Validität beschäftigen, müssen wir uns fragen, warum wir überhaupt versuchen, kausale Zusammenhänge zu schätzen, und wozu sie unserer Meinung nach dienen. In diesem Beitrag diskutieren wir, was wir unter interner und externer Validität verstehen, und erläutern die Annahmen, die erfüllt sein müssen, damit kausale Schätzungen beide Arten der Validität aufweisen. Wir betrachten Beispiele aus der soziologischen Forschung und illustrieren daran die Herausforderungen für die externe Validität. Wir appellieren an Forschende in der Soziologie, insbesondere diejenigen, die nichtexperimentelle Forschung betreiben, der externen Validität mehr Aufmerksamkeit zu schenken. Wir betonen, dass dies nicht nur für kausale Fragestellungen wichtig ist, und veranschaulichen Probleme der externen Validität, die in unterschiedlichsten nichtkausalen Studien auftreten können. Wir schließen mit praktischen Vorschlägen und Hinweisen zu einigen allgemeinen Fragen, die sich aus unserer Arbeit ergeben.

Schlüsselwörter Relevanz · Verallgemeinbarkeit · Qualitative und quantitative soziologische Forschung · Kausalität · Interne Validität

1 Introduction

Most sociologists pay little attention to external validity. This is surprising because, viewed in its broadest terms, external validity goes to the heart of what we are trying to achieve when we engage in empirical research. It is concerned with whether, and to what extent, the findings of a study apply to contexts different from the one in which they were obtained. Traditionally, the problem of external validity has been considered in relation to causal estimates derived from randomised controlled trials (RCTs), though, in fact, it applies to any claim drawn from empirical research, whether causal or descriptive, quantitative or qualitative (for discussions that concern qualitative sociology, see Hammersley 1992 and Small 2009).

In slightly more formal terms, let θ^S be a result extracted from information on a sample of units, S. S could consist of a set of responses to a questionnaire, a set of documents, a set of observations made during fieldwork, and so on. θ^S could be qualitative, such as a statement or set of statements, or quantitative, such as a mean or a covariance. It may be something that is considered descriptive or causal, such as an average treatment effect. Let θ^T be the corresponding thing in a target population that is not necessarily the population from which the sample of units was drawn.

We do not know θ^T, and the question that external validity addresses is whether, and under what circumstances, θ^S can be taken as a "good" estimate of θ^T.

In quantitative research, a classic case with which many readers will be familiar occurs when S is a random sample drawn from a population, $\mathscr{P}$, and θ^S is a parameter that we consider to be an estimate of θ^T, the same, but unknown, parameter in the population $\mathscr{P}$. Given certain assumptions, θ^S might be an unbiased and efficient estimate of θ^T (an ordinary least squares regression coefficient, for example), or it might be an asymptotically unbiased and consistent estimate (an instrumental variables estimate, for example). However, given a broader set of circumstances (for example, if S is a nonrandom sample from the population $\mathscr{P}$), we can ask under what conditions, and how in practice, can we learn about θ^T from θ^S? This is usually called the problem of generalisability. More ambitiously, suppose T is a different population, say $\mathscr{T}$, from the one from which the sample was drawn: Under what conditions, and how in practice, can we learn about θ^T from θ^S? This is the problem of transportability. In this paper we do not dwell on the generalisability/transportability distinction because we consider the former to be a special (usually less demanding) case of the latter. Instead, we focus on the sources of threats to external validity common to both generalisability and transportability.

We focus on external validity related to quantitative estimates of causal parameters, taking as our example the average treatment effect (ATE). The so-called causal revolution that has spread through economics and into adjoining social sciences has been concerned with developing methods by which to obtain credible causal estimates, especially in nonexperimental, observational settings. In the language of experiments, it has focussed on internal validity. Until recently, much less attention has been paid to external validity. But thinking about external validity obliges us to consider why we are trying to estimate causal relationships in the first place, and what we think they are for. In fields close to policy, such as epidemiology, there has been a great deal of concern with external validity and with the conditions under which causal estimates can reasonably be supposed to apply to a population beyond the sample on which the analysis was carried out. In addition, there is a large literature in social psychology and statistics dealing with the external validity of experiments, and there has also been a recent spate of publications on external validity by political scientists (for example, Mullinix et al. 2015; Findley et al. 2021; Egami and Hartman 2021, 2023; Huang et al. 2023). But amongst sociologists, these questions seem rarely to be considered (Otte et al. 2023).[1] It is not only that sociologists have not contributed to the methodological debate but also that, in empirical studies, one rarely finds explicit discussion of the range of applicability of the results.

Both internal and external validity are required for meaningful causal estimates. When correctly implemented, RCTs are often thought to secure internal validity (by the random assignment of persons to treatment and control groups) at the expense of external validity (for example, because the participants do not represent the target population, or the treatment or outcomes do not correspond to those in the target population). Field experiments and survey experiments (Falk and Heckman 2009)

[1] Exceptions are studies of the external validity of experimental results: for example, Bader et al. (2021) and Mullinix et al. (2015).

are attempts to increase the external validity of experiments. Observational studies are thought to suffer from the opposite problem: a lack of (or greater uncertainty about) internal validity but stronger claims to external validity.

There are circumstances in which external validity will not matter. Causal estimates, if they are internally valid, apply to the data in space and time from which they were derived, and it may be that our interest is in the situation captured by the specific data, such as when our interest is historical. This is obviously true of singular cases in which we encounter questions such as "What were the causes of the French revolution?" This is an example of a "causes of effects" question, which much of the modern writing on causality places beyond the scope of the "potential outcomes" approach to causality (Holland 1986). But an "effects of causes" approach might also be used to address singular historical questions—for example, "How did Basque terrorism in the late twentieth century affect the economic performance of the Basque region?" (Abadie and Gardeazabal 2003). In this case there would be no reason to ask about the external validity of the study (unless we wanted to extrapolate to other settings). But examples like these, in which it is sufficient for causal estimates to be of context-specific interest only, are the exception rather than the rule.

A general circumstance in which external validity will not be problematic is when the causal effect in question does not vary across individual units. But this kind of invariance is likely to be found only in some of the natural sciences and is unlikely to be true of the questions studied by social scientists.[2]

In this paper we discuss in greater detail what we mean by internal and external validity and set out the assumptions required for causal estimates to have both kinds of validity. We consider some examples from sociological research and the challenges to external validity that they illustrate. We urge sociologists, especially those engaged in nonexperimental research, to pay more attention to external validity.[3] And we reiterate that this is not only something that is important for causal research; in a section preceding our conclusion, we illustrate issues of external validity that may arise in noncausal studies. We conclude with some practical suggestions and remarks concerning some of the general issues that arise from our work.

[2] Modern causal identification theories differentiate four types of causal models, ranging from strongest to weakest: (a) mechanistic/physical models, which obtain the invariant causal effects with physical insights and usually involve sets of differential equations for parameter identification; (b) structural causal models, which allow us to answer counterfactual questions but, unlike mechanistic/physical models, do not allow us to obtain physical insights from the models, and the effects are not invariant across individuals; (c) causal graphical models based on directed acyclic graphs (DAGs), which allow us to predict how an outcome would change under changing distributions of predictors or artificial interventions but do not allow us to directly observe counterfactual results; (d) statistical models in settings where the units are independent and identically distributed, under which we cannot distinguish whether the effect is correlational or causational (see Peters et al. 2017, p. 11). Social science models in a closed system (for instance, in laboratory experimental settings) can be seen as structural causal models, while quasi-experiments from observational data, even when we identify all possible interventions and changes in distributions, are only causal graphical models.

[3] See Goldthorpe (2026, this issue) for some related arguments.

2 Internal Validity of Causal Effect Estimates

Internal validity is concerned with establishing "whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B" (Shadish et al. 2002, p. 38). In general, we try to secure internal validity either by the design of our research or by excluding all reasons why A and B should covary except that A causes B. In this section of the paper, we explain the requirements for an average causal effect to be unbiased—that is, to be internally valid.[4]

Imagine that we have a set of data Z, consisting of the outcome Y, treatment indicator A, and covariates V, and suppose that the treatment is binary. Under the counterfactual or potential outcomes approach to causality (Neyman 1990 [1923]; Rubin 1974), we can imagine that each individual or unit (denoted by an i subscript) has two possible values of Y, depending on whether they received treatment ($A = 1$) or not ($A = 0$). We write these $Y_i(1)$ and $Y_i(0)$; since a unit is either treated or not, one or other of these potential outcomes is unobserved.

We define the individual causal effect as the difference in individual potential outcomes

$$\tau_i = Y_i(1) - Y_i(0) \tag{1}$$

The ATE is

$$\tau = E(Y_i\,(1)) - E(Y_i\,(0)) \tag{2}$$

where the expectations are taken across all observations. However, Eqs. 1 and 2 are ideal since we cannot simultaneously observe the potential outcomes for an individual i under both the circumstances of the treatment $Y_i(1)$ and control $Y_i(0)$ (this is called the fundamental problem of causal identification). We might therefore use as an estimator of the ATE the difference between the averages of the observed outcomes among the treated and untreated:

$$\tau^* = E(Y_i|A_i = 1) - E(Y_i|A_i = 0) \tag{3}$$

Unfortunately, this naïve estimator of the ATE, τ^*, is a biased estimator of the ATE, τ. Morgan and Winship (2015, p. 46) show that this bias is equal to

$$\tau^* - \tau = \{E\,(Y\,(0)\,|A = 1) - E\,(Y\,(0)\,|A = 0)\} + \{(1-\rho)(ATT - ATU)\} \tag{4}$$

(here we have dropped the i subscript for convenience); $\rho$is the proportion of units that received treatment, ATT is the average treatment effect among those who re-

[4] We draw a distinction between unbiasedness of a causal effect, meaning that the effect is identified in the data (the assumptions required for this are given below), and statistical unbiasedness, which applies to the estimate of that effect. We usually require that statistical estimates be unbiased or consistent. In this paper, since we do not address the issue of estimation, we use unbiasedness in the former sense.

ceived treatment, and ATU is the average treatment effect among those who did not receive treatment:

$$ATT = E(Y(1)|A = 1) - E(Y(0)|A = 1) \tag{5a}$$
$$ATU = E(Y(1)|A = 0) - E(Y(0)|A = 0) \tag{5b}$$

The bias in Eq. 4 is made up of two parts, each in brackets. The first is baseline bias: The treated and the untreated would have had different average values of Y even if none of them had been treated (this is also commonly called pretreatment selection bias). The second is differential treatment effect bias: The treated and untreated differ, on average, in how much they gain from treatment. This bias comes from the heterogeneous reactions to the treatment and so is also called posttreatment heterogeneity bias. Both these biases can run in either direction, but a typical situation is one in which those who choose, or are allocated to, treatment are positively selected on other characteristics that affect Y, and they are also positively selected in that they respond to treatment more strongly than those who are not treated would respond if they were, counter to fact, treated.

For τ^* to be an unbiased estimator of the true ATE, τ, we require both of the bias terms in Eq. 4 to be zero. In observational studies we often try to eliminate baseline and differential treatment effect biases by conditioning on V, a set of confounders of the relationship between treatment and outcome. The estimator of the ATE is then

$$\hat{\tau} = \sum_{V} [E(Y_i|A_i = 1, V = v) - E(Y_i|A_i = 0, V = v)] \tag{6}$$

The ATU and ATT are examples of conditional average treatment effects, CATE. The CATE is an average treatment effect for a group most commonly defined in terms of moderating variables. That is, CATEs capture the variation in treatment effects between groups. It can be written $E(\tau_i|X_k)$for$k = 1, \ldots K$. Here, X_k denotes the k^{th} moderating variable, of which there are K in total. We write the CATEs more compactly as $\tau(X_k)$. Variables may be both moderators and confounders. For example, age might be a confounder and a moderator, and the CATE, $\tau(age)$, would capture variation in treatment effects by age. The ATE is a weighted sum or integral of CATEs:

$$\tau = \int \tau\,(x)\, dF\,(x)\,,$$

where $F(x)$ is the distribution of the moderating X variables.

In the potential outcomes framework, three assumptions must hold for internal validity:

R1: The stable unit treatment value assumption (SUTVA) requires that the potential outcomes for a unit be independent of the treatment received by any other unit. This ensures that, given a binary treatment, each unit has only two potential outcomes, regardless of how many other, or which other, units are treated. This

rules out situations in which, for example, the more people who are treated, the smaller the returns to treatment.

R2: Positivity of treatment assignment means that for each unit, the probability of receiving each treatment is greater than zero but less than one.

R3: Unconfoundedness is the assumption that potential outcomes are independent of whether a unit was treated or not. In observational studies, this is usually assumed to hold conditional on a set of confounders, V, as in Eq. 6 above.

3 External Validity

Imagine that we have selected a random sample from a population, $\mathscr{P}$. The sample used in our analyses, denoted S, is likely to deviate from a true random sample of the population in several ways: The sampling frame used might not cover the whole population, the contact rate might bias the sample away from representativeness, refusals and dropouts (in panel surveys) will have the same consequence. External validity is about inference from S to the target population, T, which may be either the original population, $\mathscr{P}$ (generalisability), or a different population, $\mathscr{T}$ (transportability).

Let Z^S be the variable set capturing the relevant features of the units in the sample: $Z^S = (Y^S, A^S, X^S)$ where Y^S represents the outcome and A^S the treatment or exposure. The set of variables X^S are the moderators of the relationship between A^S and Y^S: They capture variation in the effect of treatment between different groups or at different values of a variable. Some moderating variables may also be confounders.

We can also consider $Z^T = (Y^T, A^T, X^T)$ for the target population. There are some important differences, however, between Z^S and Z^T. Whereas we observe the full multivariate distribution of all the variables in Z^S, this is not true for Z^T, and so we cannot derive a causal estimate directly from the latter. Y^T is not measured; rather, it is defined as the outcome we would care about if the treatment A^T were applied in the population. Equally, A^T is also likely to be defined but not measured: This is the treatment we would wish to implement. Furthermore, although all the X^T corresponding to the X^S must be measured (as we explain below), we may lack information on their joint distribution: The data we have might be restricted to their individual univariate distributions or even just their average values. Nevertheless, whether τ^S, the estimated ATE in the sample (the SATE), can be considered an externally valid estimate of τ^T (the ATE in the target population, TATE) depends on the relationships between the corresponding elements of Z^S and Z^T. Threats to external validity are usually classified into four possible sources of variation between S and T: "variations in persons, settings, treatments and outcomes" (Shadish et al. 2002, p. 21; Cronbach and Shapiro 1982). Current treatments of external validity continue to use these distinctions: For example, Egami and Hartman (2023, pp. 1073–1074) refer to these as X-validity (persons), T-validity (treatments—this would be A-validity in our notation), Y-validity (outcomes), and C-validity (settings or contexts). External validity requires validity in all four of these aspects.

Variation in persons (or, more generally, in units) refers to possible differences between S and T in the X variables; for example, not all the moderators of the causal effect that are measured in S might be available in T, or they might have been measured in a different way. Variation in treatments and outcomes concerns differences in A and Y: In S we may have measured the outcome, Y, in a way that is not the same as the outcome we care about in T. For example, we might really be interested in whether someone undertakes a particular action or not (such as entering college), whereas Y^S might be a response to a survey question concerning a person's intention to do this. Similarly, the treatment A in the sample might be defined or measured in ways different from the treatment in the target population.

Cartwright (2010) has argued that causal estimates will often depend, to a degree that varies from case to case, on a wider context that is not specified in the causal model: This is captured in the idea of variation across settings or contexts. Settings may differ geographically (do the results of a study carried out in London apply in Manchester?), or they may be temporally distant (are the results of a study using data from 1995 valid today?), or both. Variation across settings means that the mechanisms that underlie the causal effects of interest are, to some degree, different across settings. External validity across settings requires that if a given unit had been, counterfactually, located in a different setting, the effects of treatment on them would be the same (Egami and Hartman 2023, p. 1075).

4 Addressing the Four External Validity Principles

An intuitive way of addressing variation between persons in the study, S, and target, T, data is to use reweighting. Reweighting is a widely used technique. For example, "poststratification reweighting" seeks to improve the representativeness of sample data by reweighting it to match the distribution of a set of demographic variables in the population from which it was drawn. In the U.S. General Social Survey, for example, reweighting yields "weighted totals that, for each GSS cross-sectional sample, equal marginal control totals from the U.S. Census Bureau estimates for education, sex, marital status, age, region of the country, race, U.S. born status, and Hispanic origin when available" (Wells et al. 2024, p. 1). Reweighting is also employed to address many other issues of interest to social scientists (for example, DiNardo et al. 1996; Fortin et al. 2011).

As noted earlier, the ATE is a weighted sum or integral of CATEs. When addressing concerns about external validity that arise from variation between persons, we reweight using data on the mediators rather than, as in the GSS example, reweighting to match the population on demographic variables. In S we have

$$\tau^S = \int \tau\left(x^S\right) dF\left(x^S\right) \tag{7a}$$

and in T:

$$\tau^T = \int \tau\left(x^T\right) dF\left(x^T\right) \tag{7b}$$

Variation across persons means that $F(x^S) \neq F(x^T)$ and so $\tau^S \neq \tau^T$—that is, the estimated ATE from S is not valid for T. However, assuming that $\tau\left(x^S\right) = \tau\left(x^T\right)$—that is, the CATEs are invariant to the setting—we can reweight the expression for τ^S to address the variation in persons and obtain an unbiased estimate of τ^T as follows[5]:

$$\hat{\tau}^T = \tau^S \times \frac{F\left(x^T\right)}{F\left(x^S\right)} \tag{8}$$

For example, suppose a car manufacturer carries out an RCT to find out how participating in a training programme affects workers' productivity. A random sample of employees in a factory is drawn, their pretreatment productivity is measured, and they are then randomly assigned to treatment (participating in the programme) or control (not participating). After training, the posttreatment productivity of individuals in both groups is measured, and the outcome, Y, is the difference between their pre-and posttraining productivity. Because the workers in the sample were randomly assigned to treatment, internal validity should, in an ideal setting, have been secured. Suppose the average effect is found to be positive but declines with the employee's age; the car manufacturer then wonders whether introducing the same programme would also improve average productivity in another of their factories, despite the fact that the age distribution of employees here is very different than in the factory where the study was carried out. Answering this question would require that the estimated CATEs from the study be reweighted to the age distribution found in the second factory under the crucial, but untestable, assumption that the age CATEs did not differ between the factories.

Reweighting is applied to address problems of both generalisability and transportability. Pearl and Bareinboim (2014, 2018; Pearl 2015) use directed acyclic graphs (DAGs) to develop rules to determine whether a particular causal effect is transportable and, if so, the kind of reweighting that should be undertaken to achieve this. But this calls for a complete model of the data-generating mechanisms involved—something we might be sceptical of attaining. As Breen (2022, p. 284) notes, "DAGs are, ideally, derived from substantive subject matter knowledge, which, in turn, means established scientific findings and well-founded theory. In many areas of science, large bodies of such findings exist, and theories are invariant across space and time or their scope conditions are known. This is less true of the social sciences".

For Eq. 8 to yield an unbiased estimate of the ATE in the target population, it requires that the sample estimate not be externally invalid because of variation across outcomes, treatments, or settings, and we also require that the measured moderating variables are complete—that is, there are no unmeasured moderating variables in either S or T. In fact, one can write a set of assumptions under which reweighting will yield externally valid estimates; these mirror those for internal validity (Stuart

[5] Several reweighting (and other) approaches have been suggested in the literature (Degtiar and Rose 2023, pp. 511–518).

et al. 2011, p. 374; Degtiar and Rose 2023, pp. 506–507; Egami and Hartman 2023, pp. 1073–1076).

R4: The mechanisms generating individual outcomes, Y, from A and X should be the same in S and T.

R5: All the effect modifiers must be measured in S, and all effect modifiers in S must also be observed in T.

R6: CATEs are independent of whether a unit is in S or T.

Notice that R4, R5, and R6 are assumptions about SUTVA, positivity, and unconfoundedness, paralleling R1, R2, and R3, except that they are concerned with invariance between S and T rather than between the treated and untreated. They may be assumed to hold unconditionally but, more often, to hold conditional on covariates. Notice, too, that these assumptions imply validity across treatments, outcome, persons, and settings. Assumption R4 allows for differences between S and T in the treatment and outcome, but the relationship between A and Y, given X, must be invariant to whether we are in S or T. This means that, for example, Y^S might be a proxy for Y^T, but it must be a sufficiently valid indicator such that R4 holds, and likewise for A^S and A^T. Assumption R5 says that there are no unobserved moderators and that all moderators observed in S, X^S, are also observed in the target population, X^T. In our earlier car factory example, all age groups found in the sample factory must be represented in the target factory, and there should be no other moderating variables. Notice that the "no unmeasured mediators" part of R5 is not secured by randomisation even though the "no unobserved confounders" (assumption R3) is. Assumption R6 directly addresses variation over settings: CATEs are assumed to be invariant across sample and target.

Degtiar and Rose (2023, pp. 508–511) suggest a number of ways that assumptions R4 to R6 can be tested, though it is impossible to fully test them—that is to say, untested assumptions always persist, but tests can at least reduce the scope of the remaining assumptions. For example, it is relatively straightforward (given the appropriate data) to test X validity by comparing the distribution of potential moderators in the sample and the target population, but external validity still assumes that there are no remaining unobserved moderators.

5 Some Examples of Threats to External Validity

5.1 Treatment and Outcome Validity

Pager and Quillian (2005) studied discrimination in hiring in Milwaukee, Wisconsin, based on race and on whether a job seeker had a criminal record. Here we concentrate on the results concerning race. Part of Pager and Quillian's study used a vignette approach in which the researchers gave employers a short written scenario describing a job applicant. The treatment, A, was the applicant's described race (White or Black), with an individual employer exposed to a randomised value of the treatment.

The outcome, Y, was the employer's expressed likelihood of hiring the hypothetical job seeker. The results showed no difference in the outcome by race. Pager and Quillian also reported the results of an experimental audit in which young men were sent to 350 randomly selected businesses to apply for advertised jobs. Here the treatment was the job seeker's perceived race, and the outcome was measured as the number of callbacks the applicant received. In this study, there was a very large difference in outcomes between the races: 34% of White individuals without a criminal record received a callback, compared with only 14% of Black individuals without a criminal record (Pager and Quillian 2005, p. 362).

When treatments and outcomes in the sample are not identical to those in the target, they must have construct validity (Findley et al. 2021, p. 371). Pager and Quillian discuss at some length how the results from the two aspects of their study should be interpreted, but their results suggest that, as a measure of discrimination in hiring, the vignette study lacks construct validity in its treatment and outcome measures. "[A] key assumption [of vignette studies] ... is that reported hypothetical behavior is an accurate proxy for the behavior that would be observed if the respondent actually encountered the situation [described in the vignette]" (Pager and Quillian 2005, p. 358). In Pager and Quillian's study, this was not the case. Generalisability of the vignette results to the population as a measure of discrimination fails because assumption R4 does not hold; the treatment (perceived race in a hypothetical scenario) and outcome (stated intention to hire) in the vignette did not correspond to the real-world treatment and outcome.

5.2 Validity Across Units

The essential requirement here is that all the mediator variables, X, should have been measured in the sample and target. This is, we believe, what authors usually have in mind when they refer, in general terms, to the representativeness of their sample data.

Some datasets are not representative of any wider population, for example when data are scraped from websites or social media platforms. Like any convenience sample, platform-specific user populations will limit generalisability because the results might not extend beyond the specific platform context.[6] Other samples, some of which are widely used, such as the Wisconsin Longitudinal Study, may be representative of a population but not the one that is of interest. Here, reweighting may offer a solution.

One challenge to X-validity arises from the selection of cases from an initially representative sample. Very commonly, sociological studies seek to achieve internal validity by controlling for all possible confounders of the causal effect. The resulting analytic sample may then be the intersection of many requirements (such as having valid values on a wide range of variables), and this may make the sample not only no longer representative of the population from which it was drawn but also

[6] In discussing problems of external validity when using digital behavioural data, Leitgöb and Keusch (2026, this issue) also stress the challenges that arise from the selectivity of samples for which such data are available.

very difficult to reweight to regain representativeness. Sharkey and Elwert (2011), in a well-executed study in which unusually close attention was given to internal validity, provide the following description, typical of sociological practice, of how their analytical sample was derived from the Panel Study of Income Dynamics (PSID):

> "In order to be included in our sample, families must meet several criteria. First, children must be assessed in the 2002 CDS [Child Development Supplement] and have nonmissing data on measures of cognitive ability. Eligibility for the 2002 CDS was based on eligibility for the original (1997) CDS, which was restricted to PSID sample families active in the survey who had children ages 0–12 in 1997. The 1997 CDS sample comprised 3563 children, and the 2002 CDS successfully recontacted and interviewed 2907 children (Mainieri 2004). Nonmissing data from the cognitive assessments are available for 2603 children.
> Second, to measure treatment status for children, information on the census tract of residence must be available for children's families in at least one year among the three survey years before the 2002 CDS (survey years 1997, 1999, and 2001). Third, background characteristics from the child's family must be available in at least one year before the measurement of the treatment status—that is, before the 1997 survey. This information is used to predict selection into the treatment for children. Fourth, to measure the treatment in the parent's generation, at least one parent must be observed, and information on the parent's census tract of residence must be available during 'childhood,' that is, in at least one year from ages 15 to 17. Fifth, background characteristics from the parent's family must be available in at least one year before age 15. This information is used to predict selection into the treatment for parents.
> The final sample comprises 1556 parent-child pairs... by construction, our sample is not representative of the current U.S. population due to extensive immigration since the late 1960s" (Sharkey and Elwert 2011, p. 1942–1944).

Reweighting such a heavily filtered sample to make it representative of a population, such as the current U.S. population, may be infeasible, leaving the study's external validity in doubt.

Another challenge arises when researchers use natural experiments because here the causal effect may not apply to the whole sample: A local average treatment effect (LATE) would be estimated instead of an average treatment effect. For example, in fixed-effects models, the causal effect is identified only from those individuals who changed treatment status. In studies using instrumental variables, the LATE is the causal effect for "compliers"; these are those units induced to take the treatment by the instrument (in contrast to "always-takers" and "never-takers" who would, respectively, take or not take the treatment irrespective of the instrument). For example, in one of the earliest applications of natural experiments in sociology, Kirk (2009) investigated whether newly released prisoners ("parolees") would be less likely to reoffend if they did not return to their previous place of residence. Focusing on New Orleans, Louisiana, he used Hurricane Katrina as an instrument. Hurricane Katrina devastated some parts of the city, and parolees who came from those areas

were largely unable to return to their old neighbourhoods. In this study, the compliers were those parolees whose decision about where to live was determined by Hurricane Katrina. The always-takers were those who would not have returned to their old neighbourhoods anyway, and the never-takers were those who, despite the hurricane, did return to their old neighbourhoods. Kirk (2009, p. 484) found that "moving away from former geographic areas substantially lowers a parolee's likelihood of re-incarceration". This is a LATE and applies only to those who moved away because of Hurricane Katrina, and so, strictly speaking, it does not tell us whether this would hold for parolees in general. To obtain an estimate that applies to the whole population of parolees, we could seek to identify the relevant X variables among the compliers, rather than the entire analytic sample, and then reweight using the population distribution of those same variables. But compliers, always-takers, and never-takers are latent groups that are not directly observed, and so, unsurprisingly, reweighting requires further untestable assumptions to be valid (Angrist and Fernández-Val 2013; Aronow and Carnegie 2013).

The use of twins, which has become increasingly popular to identify causal effects (e.g. Karlson and Birkelund 2022), presents similar problems. Families with twins, and particularly monozygotic twins, are selected samples from the population of family types. One could reweight the results from an analysis using twins to this population, but it is likely that there are unmeasured factors in play in the sample (thus violating assumption R5): For example, the influence of one twin on the other may be different, possibly larger, than the influence of one nontwin sibling on the other, leading to lower within-twin variance in whatever treatment is being measured, and since effects in twin models are driven by within-twin variation, the effects of treatment variables, A (such as education), on the outcome will be overstated, all else equal.

5.3 Validity Across Settings

As we noted earlier, validity across settings is essentially concerned with whether the same mechanisms apply in S and T. An argument often made in favour of field experiments and survey experiments over laboratory experiments is that they can more closely mirror the real-world setting to which we want our estimates to apply, and we can therefore be more confident that the assumption about an invariant mechanism holds. As a general rule, researchers should consider how the mechanisms in play in the context of their study and in the population to which they want to generalise might differ.

Temporal variation is an obvious problem of validity across settings. For example, estimates of the earnings returns to higher education must be based on data about adults, raising the question of whether such estimates can be a useful guide for those cohorts who are making the decision about whether to invest in higher education.

One particular case in which validity across settings may be threatened concerns the SUTVA assumption. For internal validity, SUTVA requires that the causal effect of treatment for each unit be independent of the treatment status of all other units. This is often an implausible assumption for questions in social science, and so assumption R1 would be violated. In studies of social networks, for example,

connections between nodes often lead to diffusion such that a treatment applied to one node may affect another, untreated node. Methods to address these issues usually require multilevel designs such that SUTVA holds between clusters (or level-2 units), even though it does not hold between level-1 units within them (see, for example, Hudgens and Halloran 2008; Sinclair et al. 2012).

Note that SUTVA can hold in the sample but not in the target population, thus violating assumption R4. One reason for this is that an intervention on a large scale may induce effects that are not present on a smaller scale. This can occur when the aggregation from the individual to the aggregate is nonlinear, that is, the whole is not just the sum of the parts. Examples of such violations of SUTVA include the dependence, all else equal, of the return to individuals of a college education on how many others graduate from college. The existence of aggregate outcomes that are not the simple aggregation of individual ones has been known to sociologists at least since the work of Schelling (1978). This is likely to be a particular problem when policy recommendations are made. Consider, for example, an experiment (such as Krueger 1999) or natural experiment (such as Angrist and Lavy 1999) that finds that classroom size affects academic performance. Extrapolating the results to infer the likely consequences of a policy to reduce classroom sizes throughout a country or region is fraught with difficulties. In particular, parents may respond to the reduction in class size by reducing their other educational investments in their children: In this case, the returns to the policy would be less than the original research had indicated. Another possibility is that parents of children who had been advantaged by being in smaller classrooms may seek to preserve their children's advantages in other ways (such as investing in tutoring or other forms of shadow education). In this case, the policy might bring about larger than anticipated overall gains but would leave educational inequality unchanged.

6 External Validity Applies to All Kinds of Sociological Research

Although our exposition has mostly dealt with causal analyses, we want to reiterate that external validity should also matter for descriptive studies and for qualitative research. The failure of much qualitive research, especially ethnography, to be generalisable has been acknowledged by some of its practitioners (Hammersley 1992, especially Chap. 5; Small 2009) as well as its critics (Goldthorpe 2000, Chap. 4). Problems of both internal and external validity have been raised. In relation to the former, the main issue is whether the data obtained are representative of the context, locale, or group under study (due, in some ethnographies, to a reliance on a small number of key informants). The latter concerns whether the findings that apply to the context, locale, or group are generalisable to other settings.

Validity across settings is a potential problem for most kinds of social science research. Social mobility research is primarily descriptive, rather than causal, and although it is often comparative (for example, Breen and Müller 2020), temporal variation is an obvious source of external invalidity. Such studies are, of necessity, retrospective in that they require information on adults who are old enough to have reached what is sometimes called "occupational maturity". They must therefore

have experienced the mobility process, including acquiring an education and a job and possibly experiencing career mobility. Estimates of social mobility are typically presented for the population aged between 30 and 60 years. Whether such estimates are informative about the mobility that later cohorts will experience depends on whether the mechanisms that drive mobility are invariant across cohorts. Since these mechanisms are not known fully, this question is difficult to answer.

An example in which geographic variation challenges external validity is research into the relationship between grandparents and grandchildren. Often the issue in question is whether there is an association between the same outcome or characteristic of grandparents and grandchildren, net of the characteristic of the grandchildren's parents. Examples of this design using social class include the work of Erola and Moisio (2007) and Chan and Boliver (2013). Other studies focus on other outcomes, with educational attainment being a popular choice (e.g. Warren and Hauser 1997; Ziefle 2016). To take a rather extreme example, a Nordic study might find that grandparents' education was independent of grandchildren's, net of parents', education. But we would not expect this finding to be valid for East Asian countries, where different social structures, family organisations, and cultural traditions lead to grandparents having a powerful bearing on their grandchildren.[7] The differences between the two kinds of society are not captured by differences in the distribution of what a causal analysis would call invariance of conditional average effects across settings; rather, the contextual differences are such that the invariance assumption itself is not plausible. More simply, the differences reside in different mechanisms. Understanding these mechanisms in both the sample and the target population is necessary for any assessment of external validity.

External validity depends not only on the validity of the measures used in the research (specifically, A- and Y-validity, as we saw in our discussion of Pager and Quillian 2005) but also on the representativeness of our data and, in some cases, the processes by which it was obtained. The latter is particularly important for research, such as vignette and other experimental or quasi-experimental studies, where we need to consider the extent to which the processes in the study mirror those in the real world that we want to say something about. For example, vignette studies of hiring decisions implicitly assume that whether to hire or not, or whether to call back an applicant or not, is an individual decision, whereas in the real world, these decisions are often made by a group of people, such as a hiring committee, during whose deliberations individual decisions may be discounted or may change. A further aspect of representativeness is, of course, the composition of the sample: Are the survey participants representative of the kind of people who make hiring decisions in real companies and businesses (for example, Marquis et al. 2024, pp. 1743–1744)? If they are not, then even if we had secured A- and Y-validity, the study's results could not be considered externally valid.

A distinction in experimental research in social science is sometimes made between empirically driven and theory-driven experiments. The latter are usually laboratory or field experiments that seek to test an explanation or mechanism derived

[7] For a historical example, see Song et al. (2015). For a contemporary example, see Zeng and Xie (2014). Song (2021) discusses some relevant methodological concerns.

from previous research or theory. What is being tested in such cases is not an empirical claim but a theoretical construct (Gërxhani and Miller 2022, pp. 314; Deaton and Cartwright 2018) that, if supported, may be applied to other, broader contexts to explain empirical results. For experiments of this sort, it is argued, external validity will not be a cause of concern. Many observational studies also claim to focus on testing hypotheses or mechanisms (including two of the studies referred to above, those of Kirk 2009 and Sharkey and Elwert 2011), so is it true that issues of external validity can safely be ignored?[8] We believe not if we care about the scope of the mechanism that is being tested: that is to say, whether the mechanism is invariant over some range of settings or in a target population (see Goldthorpe 2000, pp. 81–83, for a discussion of this question in the context of ethnographic studies). Here, replication would play a crucial role: If the same mechanism were found to operate in a variety of settings, possibly using different analytical designs and methods, we might have more confidence in generalising to a wider population. Nevertheless, in these studies a mechanism is tested through a quantitative analysis leading to the question "Would we get the same quantitative result (implying the same mechanism) in the target population?" Here we might be concerned less with the numerical value of the results and more with its sign (Egami and Hartman 2023), but we would still have to consider the four aspects of external validity that we have presented here.

7 Some Guidelines

The first question to be addressed in considering the external validity of a study is to what population or group the researcher wants to generalise: Findings cannot be said to be externally valid without specifying the target. The extent to which studies omit this information would be surprising if it were not so common. As Hammersley (1992, p. 87), writing about ethnographies, points out, "ethnographers sometimes write as if they are generalising to a category of phenomena occurring in unspecified times and places, rather than to an identified aggregate of settings during a specific time period". This criticism might, with equal justification, be levelled at most others forms of sociological research.[9] In many (perhaps most) studies, the population to which the results are meant to generalise is left unstated, though implicitly it may be the population from which the sample was drawn. Yet, as we showed earlier, even if the probabilities of selection into the original sample were known, how the data were processed to arrive at an analytical sample call this into question. Thus, stating to whom the results are meant to apply should be a feature of all sociological research.

The next step would be to consider the plausibility of the specified claims to external validity in terms of the four dimensions discussed above.[10] Here, DAGs

[8] Scepticism is in order here, not least because of the heterogeneity of causal effects.

[9] And perhaps particularly those dealing with the United States, where findings are sometimes presented as though they were of universal application.

[10] Findley et al. (2021, pp. 376–383) deal with this in some detail in a section of their paper that seeks to develop "evaluative criteria for external validity".

may be helpful insofar as they faithfully represent the data-generating process for both the sample and the target. Their use will also make clear which, and how far, elements of external validity can be assessed empirically and how far they rely on assumptions. In the latter case, these assumptions should be justified in a similar way to which, for example, the untestable assumptions that are required for the internal validity are discussed (typically for instrumental variable or other quasi-experimental designs). For a good sociological example of how assumptions necessary for a finding to have a causal interpretation can be justified, see Torche (2011). Important untestable assumptions for external validity are that there are no unobserved moderators and that the CATEs are invariant to context (R5 and R6). Arguments about whether or not these are plausible should ideally rely on well-founded theory and subject matter knowledge (as in our example of grandparental effects in Europe and East Asia).

An ideal study would be both internally and externally valid, but, in practice, there may be a trade-off between them. If, in a study, an instrumental variable (IV) were available, the trade-off would be between a consistent and internally valid IV estimate of the treatment effect that applied only to compliers and an estimate using a method that controlled for only observed confounders (a multiple regression or a weighting approach, for example) and that would therefore not be internally valid but would apply to the sample as a whole and might therefore be easier to reweight to the target population.[11]

In recent years, more attention has begun to be paid to the importance of replication. Replication and external validity are, of course, different issues, but thinking about them together raises the question of whether, or in what circumstances, it makes sense to talk about generalising or transporting the results from a single study, given that the results from any single study are likely to be contaminated with some degree of error. Given a clear delineation of the target population to which the results of a study should generalise, replication, using samples or other data whose results should generalise to the same target population, could be very useful in establishing the findings that are indeed generalisable or transportable (Otte et al. 2023).[12]

8 Conclusions

The main goal of this paper has been to make sociologists more sensitive to the issue of external validity, and we have therefore sought to present an accessible summary of contemporary approaches to it. The dearth of studies in sociology that address the problem has meant that we have largely drawn on work in epidemiology, statistics, and political science. This overwhelmingly deals with the external validity

[11] Because ordinary least squares estimates usually have smaller standard errors than IV estimates, the trade-off between the two is conventionally considered in terms of bias and variance, but with external validity in mind, the broader trade-off would be between bias and both variance and generalisability.

[12] Ideally, researchers should design multisite studies or follow-up studies in new settings as part of a research program.

of estimates from experiments that often have a close link to policy. But to argue that this literature does not apply to sociology would, we believe, be a mistake. Many sociological studies that seek to estimate a causal effect from observational data do claim to be relevant for policy. More important, almost every piece of empirical sociological research makes claims that extend beyond the particular data used in the study. Such claims cannot be substantiated without considerations of external validity. Bringing external validity into sociological research means routinely asking *"To whom and where do our results apply?"* and addressing that question with as much care as we devote to internal validity.

Acknowledgements The authors thank John Ermisch, Said Hassan, Steffen Hillmert, and the editors and contributors to this volume for helpful comments.

Conflict of interest R. Breen and G. Pan declare that they have no competing interests.

References

Abadie, Alberto, and Javier Gardeazabal. 2003. The economic costs of conflict: A case study of the Basque country. *American Economic Review* 93: 113–32.

Angrist, Joshua, and Iván Fernández-Val. 2013. ExtrapoLATE-ing: External validity and overidentification in the LATE framework. In *Advances in economics and econometrics: Tenth world congress*, eds. Daren Acemoglu, Manuel Arellano, and Eddie Dekel, 401–443. Cambridge: Cambridge University Press.

Angrist, Joshua, and Victor Lavy. 1999. Using Maimonides' Rule to estimate the effect of class size on scholastic Achievement. *Quarterly Journal of Economics* 114: 533–575.

Aronow, Peter M., and Allison Carnegie. 2013. Beyond LATE: Estimation of the average treatment effect with an instrumental variable. *Political Analysis* 21: 492–506.

Bader, Felix, Bastian Baumeister, Roger Berger, and Marc Keuschnigg. 2021. On the transportability of laboratory results. *Sociological Methods & Research* 50: 1452–1481.

Breen, Richard. 2022. Causal inference with observational data. In *Handbook of sociological science: Contributions to rigorous sociology*, eds. Klarita Gërxhani, Nan D. de Graaf, and Werner Raub, 272–286. Cheltenham: Edward Elgar.

Breen, Richard, and Walter Müller (eds.). 2020. *Education and intergenerational social mobility in Europe and the United States*. Stanford, CA: Stanford University Press.

Cartwright, Nancy. 2010. What are randomized control trials good for? *Philosophical Studies* 147: 59–70.

Chan, Tak Wing, and Vikki Boliver. 2013. The grandparents effect in social mobility: Evidence from British birth cohort studies. *American Sociological Review* 78: 662–678.

Cronbach, Lee J., and Karen Shapiro. 1982. *Designing evaluations of educational and social programs*. San Francisco, CA: Jossey Bass.

Deaton, Angus, and Nancy Cartwright. 2018. Understanding and misunderstanding randomized control trials. *Social Science and Medicine* 210: 2–21.

Degtiar, Irina, and Sherri Rose. 2023. A review of generalizability and transportability. *Annual Review of Statistics and Its Application* 10: 501–524.
DiNardo, John, Nicole M. Fortin, and Thomas Lemieux. 1996. Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach. *Econometrica* 64: 1001–1044.
Egami, Naoki, and Erin Hartman. 2021. Covariate selection for generalizing experimental results: Application to a large-scale development program in Uganda. *Journal of the Royal Statistical Society Series A* 184: 1524–1548.
Egami, Naoki, and Erin Hartman. 2023. Elements of external validity: Framework, design and analysis. *American Political Science Review* 117: 1070–1088.
Erola, Jani, and Pasi Moisio. 2007. Social mobility over three generations in Finland, 1950–2000. *European Sociological Review* 23: 169–183.
Falk, Armin, and James J. Heckman. 2009. Lab experiments are a major source of knowledge in the social sciences. *Science* 326: 535–538.
Findley, Michael G., Kyosuke Kikuta, and Michael Denly. 2021. External validity. *Annual Review of Political Science* 24: 365–393.
Fortin, Nicole, Thomas Lemieux, and Sergio Firpo. 2011. Decomposition methods in economics. In *Handbook of labor economics volume 4, Part A*, eds. Orley Ashenfelter, and David Vard, 1–102. Amsterdam: Elsevier.
Gërxhani, Klarita, and Luis Miller. 2022. Experimental sociology. In *Handbook of sociological science*, eds. Klarita Gërxhani, Nan D. de Graaf, and Werner Raub, 309–323. Cheltenham: Edward Elgar.
Goldthorpe, John H. 2000. *On sociology: Numbers, narratives, and the integration of research and theory*. Oxford: Oxford University Press.
Goldthorpe, John H. 2026. Description, causal explanation and policy intervention in sociology. *This issue.*
Hammersley, Martyn. 1992. *What's wrong with ethnography? Methodological explanations*. London: Routledge.
Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81: 945–960.
Huang, Melody, Naoki Egami, Erin Hartman, and Luke Miratrix. 2023. Leveraging population outcomes to improve the generalization of experimental results: Application to the JTPA study. *The Annals of Applied Statistics* 17: 2139–2164.
Hudgens, Michael G., and M. Elizabeth Halloran. 2008. Toward Causal Inference with Interference. *Journal of the American Statistical Association* 103: 832–842.
Karlson, Kristian B., and Jesper F. Birkelund. 2022. Family background, educational qualifications, and labour market attainment: Evidence from Danish siblings. *European Sociological Review* 38: 988–1000.
Kirk, David S. 2009. A natural experiment on residential change and recidivism: Lessons from hurricane Katrina. *American Sociological Review* 74: 484–505.
Krueger, Alan B. 1999. Experimental estimates of education production functions. *Quarterly Journal of Economics* 114: 497–532.
Leitgöb, Heinz, and Florian Keusch. 2026. Causal inferences from digital behavioral data: Methodological implications. *This issue.*
Marquis, Christoper, András Tilcsik, and Ying Zhang. 2024. Attractiveness and attainment: Status, beauty, and jobs in China and the United States. *American Journal of Sociology* 129: 1720–1762.
Morgan, Stephen L., and Chris Winship. 2015. *Counterfactuals and causal inference: Methods and principles for social research. Second edition.* New York: Cambridge University Press.
Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2015. The generalizability of survey experiments. *Journal of Experimental Political Science* 2: 109–138.
Neman, Jerzy S. 1990. On The Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science* 5: 465–472. Translated from the 1923 Polish original by Dorota M. Dabrowska and Terence P. Speed.
Otte, Gunnar, Tim Sawert, Josef Brüderl, Stefanie Kley, Clemens Kroneberg, and Ingo Rohlfing. 2023. Gütekriterien in der Soziologie. Eine analytisch-empirische Perspektive. *Zeitschrift für Soziologie* 52: 26–49.
Pager, Devah, and Lincoln Quillian. 2005. Walking the talk? What employers say versus what they do. *American Sociological Review* 70: 355–380.
Pearl, Judea. 2015. Generalizing experimental findings. *Journal of Causal Inference* 3: 259–266.
Pearl, Judea, and Elias Bareinboim. 2014. External validity: From do-calculus to transportability across populations. *Statistical Science* 29: 579–595.

Pearl, Judea, and Elias Barenboim. 2018. *Transportability across studies: A formal approach*. Los Angeles: Computer Science Department, University of California.
Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: Foundations and learning algorithms*. Cambridge, MA: MIT Press.
Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66: 688–701.
Schelling, Thomas C. 1978. *Micromotives and macrobehavior*. New York: Norton.
Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.
Sharkey, Patrick, and Felix Elwert. 2011. The legacy of disadvantage: Multigenerational neighbourhood effects on cognitive ability. *American Journal of Sociology* 116: 1934–1981.
Sinclair, Betsy, Margaret McConnell, and Donald P. Green. 2012. Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science* 56: 1055–1069.
Small, Mario L. 2009. 'How many cases do I need?' On science and the logic of case selection in field-based research. *Ethnography* 10: 5–38.
Song, Xi. 2021. Multigenerational mobility: A demographic approach. *Sociological Methodology* 51: 1–43.
Song, Xi, Cameron D. Campbell, and James Z. Lee. 2015. Ancestry matters: Patrilineage growth and extinction. *American Sociological Review* 80: 574–602.
Stuart, Elizabeth A., Stephen R. Cole, Catherine P. Bradshaw, and Philip J. Leaf. 2011. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A* 174: 369–386.
Torche, Florencia. 2011. The effect of maternal stress on birth outcomes: Exploiting a natural experiment. *Demography* 48: 1473–1491.
Warren, John R., and Robert M. Hauser. 1997. Social stratification across three Generations: New evidence from the Wisconsin Longitudinal Survey. *American Sociological Review* 62: 561–572.
Wells, Brian M., Zachary H. Seeskin, and Amy Ihde. 2024. Post-stratification Weights for GSS 1972–2022. GSS Methodological Report 137. *NORC*. https://gss.norc.org/content/dam/gss/get-documentation/pdf/reports/methodological-reports/GSS%20MR137%20Poststratification%20Weights.pdf.
Zeng, Zhen, and Yu Xie. 2014. The effects of grandparents on children's schooling: Evidence from rural China. *Demography* 51: 599–617.
Ziefle, Andrea. 2016. Persistent educational advantage across three generations: Empirical evidence for germany. *Sociological Science* 3: 1077–1102.

Richard Breen is Emeritus Fellow of Nuffield College, University of Oxford. His research interests are inequality, intergenerational mobility, and quantitative methods. Recent publications have appeared in *Sociological Science* and *The European Sociological Review*. Together with Walter Müller he edited *Education and Intergenerational Social Mobility in Europe and the United States* (Stanford University Press, 2020). He is a Fellow of the British Academy, a Fellow of the European Academy of Sociology and a Member of the Royal Irish Academy and Academia Europaea.

Guanghui Pan is a DPhil student in Sociology at the University of Oxford. His research focuses on social stratification and mobility, as well as quantitative methods, especially causal inference and doubly robust machine-learning approaches. His DPhil thesis applies doubly robust, efficient estimators to causal inference with time-varying social variables.