

Natural Experiments: Missed Opportunities for Causal Inference in Psychology



Michael P. Grosz¹, Adam Ayaita², Ruben C. Arslan^{3,4},
Susanne Buecker⁵, Tobias Ebert^{6,7}, Paul Hünermund⁸,
Sandrine R. Müller⁹, Sven Rieger¹⁰, Alexandra Zapko-Willmes¹¹,
and Julia M. Rohrer³

¹Institute for Mind, Brain and Behavior, HMU Health and Medical University Potsdam, Potsdam, Germany; ²School of Business and Economics, RWTH Aachen University, Aachen, Germany; ³Wilhelm Wundt Institute for Psychology, Leipzig University, Leipzig, Germany; ⁴Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany; ⁵Institute of Psychology, German Sport University Cologne, Cologne, Germany; ⁶School of Social Sciences, University of Mannheim, Mannheim, Germany; ⁷Institute of Behavioral Science and Technology, University of St. Gallen, St. Gallen, Switzerland; ⁸Department of Strategy and Innovation, Copenhagen Business School, Frederiksberg, Denmark; ⁹Google LLC, New York, New York; ¹⁰Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany; and ¹¹Department of Psychology, University of Bremen, Bremen, Germany

Advances in Methods and Practices in Psychological Science
January-March 2024, Vol. 7, No. 1,
pp. 1–15
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/25152459231218610
www.psychologicalscience.org/AMPPS



Abstract

Knowledge about causal effects is essential for building useful theories and designing effective interventions. The preferred design for learning about causal effects is randomized experiments (i.e., studies in which the researchers randomly assign units to treatment and control conditions). However, randomized experiments are often unethical or unfeasible. On the other hand, observational studies are usually feasible but lack the random assignment that renders randomized experiments causally informative. Natural experiments can sometimes offer unique opportunities for dealing with this dilemma, allowing causal inference on the basis of events that are not controlled by researchers but that nevertheless establish random or as-if random assignment to treatment and control conditions. Yet psychological researchers have rarely exploited natural experiments. To remedy this shortage, we describe three main types of studies exploiting natural experiments (standard natural experiments, instrumental-variable designs, and regression-discontinuity designs) and provide examples from psychology and economics to illustrate how natural experiments can be harnessed. Natural experiments are challenging to find, provide information about only specific causal effects, and involve assumptions that are difficult to validate empirically. Nevertheless, we argue that natural experiments provide valuable causal-inference opportunities that have not yet been sufficiently exploited by psychologists.

Keywords

causality, nonexperimental, regression discontinuity design, instrumental variable estimation, observational studies

Received 12/21/22; Revision accepted 11/1/23

Knowledge about causal effects is essential for building useful theories and designing effective interventions. One method for learning about causal effects is randomized experiments, that is, studies in which researchers randomly assign units (e.g., individuals, classrooms, hospitals) to treatment and control conditions. Randomized experiments are the preferred way of learning about causal effects because the random assignment eliminates many alternative explanations for an apparent treatment

effect (e.g., Murnane & Willet, 2011; Pearl, 2009; Shadish et al., 2002).

In randomized laboratory experiments, researchers have a high level of control not only over the assignment

Corresponding Author:

Michael P. Grosz, Institute for Mind, Brain and Behavior, HMU Health and Medical University Potsdam
Email: michael.grosz@health-and-medical-university.de



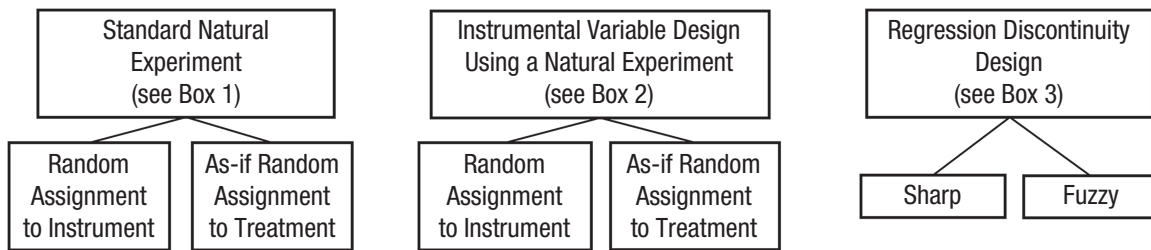


Fig. 1. Types of studies exploiting natural experiments. We distinguish between three types of studies exploiting natural experiments as implemented by Dunning (2012) and Sieweke and Santoni (2020).

of units but also over experimental conditions. This might enable researchers to demonstrate a causal effect even when there are hardly any opportunities to do so in the field. However, it is often unfeasible or unethical to create the conditions of interest in a randomized experiment. If it is feasible and ethical, the creation of the conditions by researchers may shape participants' beliefs about researchers' intentions or hypotheses or expectations more generally, which might alter participants' behavior, a phenomenon called "demand effects" (e.g., Corneille & Lush, 2023). Finally, the effect sizes found in randomized experiments—particularly of those conducted in a laboratory setting—might not generalize to the populations, contexts, and conditions of interest because the samples, study conditions, and experimental manipulations are not representative of the populations of interest or of conditions, events, or interventions that occur in real life (e.g., Cesario, 2022; Diener et al., 2022; Galizzi & Navarro-Martinez, 2019; Shadish et al., 2002).

These issues tend to affect observational studies to a smaller extent. For example, data may be collected as part of large-scale survey studies without any immediately apparent study goal, thus avoiding certain types of demand effects. And because participants receive the "treatment" in real-life conditions, the generalizability to the context of interest is often higher than in laboratory experiments. However, all of these benefits come at a high price: Because of the lack of random assignment, it is often impossible to rule out that differences between the treated and untreated conditions are the result of confounding factors (e.g., Schafer & Kang, 2008).

Natural experiments offer unique opportunities to combine features of randomized experiments and observational studies. A natural experiment is a "naturally" occurring event or condition (i.e., an event or condition not created by researchers) that affects some but not all units of a population (e.g., Dunning, 2012; Sieweke & Santoni, 2020). What sets natural experiments apart from events and conditions that are studied in standard observational studies is that people are randomly or as-if randomly assigned to treatment and control conditions. This (as-if) random assignment also sets natural experiments apart from quasi-experiments, at least according

to many common definitions of quasi-experiments.¹ Hence, in natural experiments, it can be assumed that there are no, or almost no, systematic differences between the treated and untreated individuals before the treatment. A classic example is the Vietnam lottery draft, in which a lottery determined which men were called to military service in the Vietnam War (e.g., Angrist, 1990). Natural experiments differ from (non-natural) randomized experiments in that participants are not randomly assigned to the treatment and control groups by researchers, and researchers do not control the experimental manipulation and conditions. Hence, "natural experiments are not so much designed as discovered" (Dunning, 2012, p. 41).

The current work intends to promote natural experiments in the field of psychology, an area in which they have hardly been used so far, as we demonstrate in the following section. We also clarify the advantages of natural experiments, describe several types of natural experiments (for an overview, see Fig. 1), and provide inspiring examples that illustrate the potential of natural experiments in psychology. We want to familiarize readers with the general idea underlying these designs without too many technical details. We provide empirical examples with directed acyclic graphs² to illustrate some analysis options in Boxes 1 to 3. Excellent, more technical introductions to natural experiments, as well as discussions of their potential pitfalls, are available elsewhere and should be consulted when readers have decided to use natural experiments in their own work (for key references, see Boxes 1–3).

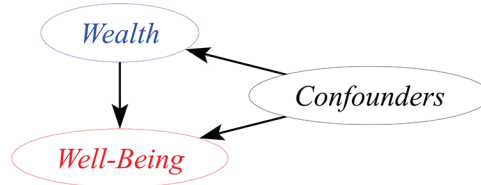
How Frequently Are Natural Experiments Used in Psychological Research?

Psychological researchers rarely mention or make use of natural experiments. We conducted an electronic database search. It indicated that natural experiments were mentioned in 0.07% of all psychology abstracts that mention the word "study," "studies," "experiment," or "data," whereas natural experiments were mentioned in 1.50% of all business and economics abstracts that mention the word "study," "studies," "experiment," or "data."³ We

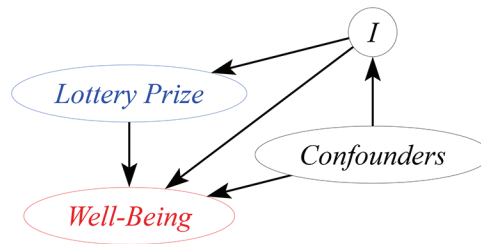
Box 1. Standard Natural Experiment

Example article: Lindqvist et al. (2020)

Initial identification challenge: Wealth and well-being may be confounded by common causes such as education, fluid intelligence, and conscientiousness.



(As-if) random assignment exploited to identify effect: In a large administrative sample of lottery participants, lottery prizes are randomly assigned conditional on certain factors (e.g., number of lottery tickets). **I** is a matrix of indicator variables reflecting groups of lottery players within which prize money can be considered randomly assigned. Each group consisted of a large-prize winner and controls that were exactly matched on number of lottery tickets (in the month of win) and other variables.



Typical analysis model: Fixed-effects linear regression:

$$Y_i = \alpha X_i + C\beta + I\gamma + e_i$$

Y_i = measure of well-being standardized to unit variance for respondent i measured after the lottery event

X_i = lottery prize (in \$100,000)

C = matrix of baseline characteristics measured before the lottery event (year of birth, sex, college degree, Swedish-born, married, number of children, capital income, labor income)

I = matrix of indicator variables (dummy variables) for groups of lottery players within which prize money is randomly assigned

e_i = error term

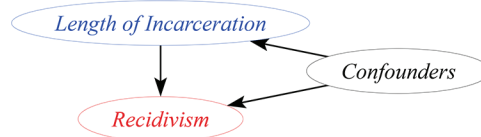
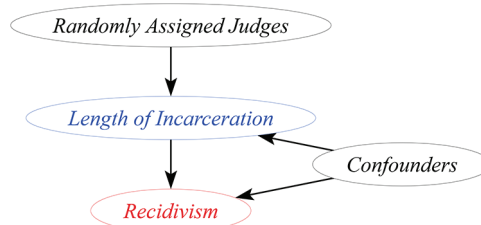
Annotated key references for further information: Dunning (2012) elaborate on the types of studies using natural experiments and their analysis and strengths and weaknesses, with many examples from political sciences and economics. Sieweke & Santoni (2020) provide guidelines for how to discover and analyze natural experiments (including robustness checks); the authors include a systematic review of 87 studies that used natural experiments in leadership research.

Note: For didactic reasons, Box 1 presents a slightly modified version of the design, notation, and analysis reported in Lindqvist et al. (2020). Red circles = outcome variable; blue circles = treatment variable.

compared psychology to business and economics because economics is known for its focus on causal inference. Other related fields such as sociology or political science presumably fall somewhere in the middle ground between psychology and business and economics in terms of natural experiments. For example, natural experiments were mentioned in 0.32% of all sociology abstracts that mention the word “study,” “studies,” “experiment,” or “data.” Furthermore, we systematically reviewed and coded the research designs of 216

randomly sampled articles published in psychology and economics flagship journals in 2019. None of the 108 reviewed psychology articles used a natural experiment, whereas 36 of the 108 reviewed economics articles used a natural experiment (for details, see Table 1).

Reasons for the more frequent use of natural experiments in economics than psychology might be that randomized experiments are hardly feasible in macro-economics because researchers cannot experiment with

Box 2. Instrumental-Variable Estimation Using a Natural Experiment**Example article:** Green & Winik (2010)**Initial identification challenge:** Length of incarceration and recidivism may be confounded by common causes such as the (home) environment or the personality of the offender.**(As-if) random assignment exploited to identify effect:** Drug offenders were as-if randomly assigned to one of nine different judicial calendars and thus to judges who differed in terms of their punitiveness/leniency.**Typical analysis model.** Two-stage least-squares regression:

Stage 1:

$$X_i = s_0 + s_1 Z_{1i} + s_2 Z_{2i} + \dots + s_8 Z_{8i} + \mathbf{C}\beta + e_i$$

 X_i = length of incarceration Z_{1i} = indicator variable (dummy variable) indicating whether offender was assigned to Judicial Calendar 1 or not \mathbf{C} = matrix of control variables that were determined before judge assignment: demographics, prior record variables, charge variables, and drug type e_i = error term

Stage 2:

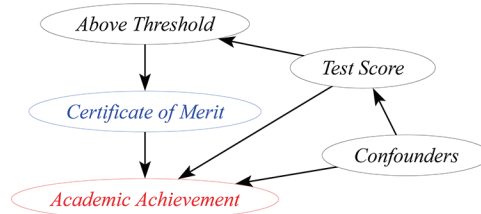
$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \mathbf{CT} + u_i$$

 Y_i = recidivism \hat{X}_i = predicted incarceration (from Stage 1) u_i = error term

In Stage 2, the standard errors need to be corrected because performing an ordinary least-squares regression using the predicted values as predictors will yield incorrect standard errors (e.g., Angrist & Pischke, 2009).

Annotated key references for further information: Bastardo et al. (2023) discuss conditions that valid instruments must satisfy, conduct data simulations to demonstrate the sensitivity to violations of their conditions, provide descriptions of common mistakes, and offer nontechnical guidelines targeted at the study design, analysis, and reporting phases. Morgan & Winship (2015) provide (among other things) an accessible introduction to instrumental-variable estimation, with an emphasis on effect heterogeneity. Brito & Pearl (2002) provide precise graphical conditions for the validity of generalized instrumental variables in directed acyclic graphs.

Note: For didactic reasons, Box 2 presents a slightly modified version of the design, notation, and analysis reported in Green and Winik (2010). The data and R code for the analysis reported in Green and Winik can be found at <https://isps.yale.edu/research/data/d028>. Red circles = outcome variable; blue circles = treatment variable.

Box 3. Regression-Discontinuity Design**Example article:** Thistlethwaite & Campbell (1960)**Initial identification challenge:** Students above a certain threshold of test scores receive a certificate of merit. Scoring above the threshold and subsequent academic achievement may be confounded by common causes such as intelligence, personality, and social support.**Continuity of confounder distribution exploited to identify effect:** Distribution of confounders is assumed to be a smooth function around the threshold. Therefore, differences between students who fall just below and above the test score threshold are regarded as negligible, and confounding can thus be ignored.**Typical analysis model:**

There is a global approach with parametric regression models:

$$Y_i = b_0 + b_1 Z_i + b_2 Z_i^2 + b_p Z_i^p + \gamma X_i + e$$

 Y_i = measure of subsequent academic achievement Z_i = normalized test scores, which qualify for certificate of merit when above zero; included up to polynomial of degree p to estimate a flexible function $f(Z_i)$ X_i = dummy variable indicating award of certificate of merit

Global polynomial regressions can be unstable at the discontinuity point. Therefore, local nonparametric regressions are preferable (Gelman & Imbens, 2019). In the local approach with nonparametric regression models, local linear regressions are estimated separately below and above the threshold:

$$(\hat{\alpha}, \hat{\beta}) = \operatorname{argmin}_{\alpha, \beta} \sum_{i=1}^n (Y_i - \alpha - \beta Z_i)^2 K\left(\frac{Z_i}{b}\right)$$

 α = intercept of local linear regression β = slope of local linear regression b = bandwidthLinear regressions are estimated locally for every point of Z_i and then combined to one smooth function. The bandwidth b determines how many data points around a given value of Z_i are included in the estimation. Kernel function $K()$ puts more weight on observations in a local neighborhood of Z_i . Optimal bandwidth b^* can be chosen in a data-driven way (Imbens & Kalyanaraman, 2012).**Annotated key references for further information:** Cattaneo et al. (2020) introduce the fundamentals of regression-discontinuity designs, providing pragmatic guidance for analysis, including R and Stata code.

Cattaneo et al. (2023) consider extensions such as fuzzy regression-discontinuity designs.

Huntington-Klein (2022) covers (among other things) regression-discontinuity designs, with a strong emphasis on data-generating processes and causal diagrams. The book features extensive code examples in R, Stata, and Python. Thoemmes et al. (2017) accessibly introduce and provide guidelines for the analysis of regression-discontinuity designs using R. Noticeably, they use actual empirical data to illustrate all steps of the analysis of a regression-discontinuity design.

Note: For didactic reasons, Box 3 presents a modified version of the design, notation, and analysis reported in Thistlethwaite and Campbell (1960). Red circles = outcome variable; blue circles = treatment variable.

Table 1. Review of a Random Sample of Economics and Psychology Articles

	Economics	Psychology
Number of articles reviewed	108	108
Number of articles containing . . .		
An empirical study	88	96
A randomized experiment	17	42
A natural experiment	36 ^a	0
A standard natural experiment with true randomization	1	0
A standard natural experiment with as-if randomization	19	0
An instrumental-variable design using a natural experiment with true randomization	1	0
An instrumental-variable design using a natural experiment with as-if randomization	18	0
A sharp regression-discontinuity design	3	0
A fuzzy regression-discontinuity design	2	0

Note: We reviewed a random sample of 216 articles published in eight flagship journals from psychology and economics in the year 2019. We sampled articles from four empirical psychology journals that had a relatively high impact according to the 2021 SCImago Journal Rank (*Journal of Applied Psychology*, *Journal of Personality and Social Psychology*, *Psychological Science*, and *Clinical Psychological Science*) and the four of the top five economic journals (e.g., Heckman & Moktan, 2020) that publish largely empirical studies (*American Economic Review*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*). A student assistant coded basic information (e.g., authors, year, DOI, and whether the article contains an empirical study or not). The study designs of each article were independently coded by either two authors of the current work or by one of the authors and the student assistant (for coding manual, data, and interrater agreements, see Tables S1 to S3 on the OSF at <https://osf.io/a5nxxm>). Disagreements and uncertainties were resolved by the first author. ^aOf the 36 natural experiments, 11 were borderline cases (for details, see Table S2). The total number of articles using natural experiments (36) is smaller than the sum of articles using specific types of natural experiments because some articles used more than one type.

countries' economies, rendering natural experiments such as public policies an attractive alternative to randomized experiments. Moreover, economists often use administrative observational data to exploit a natural experiment (e.g., Angrist, 1990). Administrative data rarely include psychological measures that are needed to answer psychological research questions. Nevertheless, it would be possible to purposefully collect the required variables months or years after the natural experiment occurred (e.g., Ertola Navajas et al., 2022). A further reason for the lack of usage of and awareness about natural experiments in psychology might be a difference in research culture across fields. Whereas economists have embraced the challenge of estimating causal effects on the basis of observational data, psychologists have traditionally avoided making explicit inferences about causal effects in observational studies (e.g., Grosz et al., 2020). We think this avoidance of explicit causal inference in the absence of a randomized experiment has to some degree led researchers to neglect (the concept of) natural experiments. The resulting lack of studies exploiting natural experiments in the psychology literature has further exacerbated the problem because published work often inspires other researchers to use the same methodology for their own research questions—that is, to exploit the same or to discover a different natural experiment (Dunning, 2012).

Why Psychologists Should Use Natural Experiments

We believe psychologists should use natural experiments more often than they currently do for several reasons. Because of the natural occurrence of the treatment in natural experiments, they can be an option when randomized experiments are unethical or unfeasible. For example, it is unethical and unfeasible to experimentally induce an earthquake to study its effects, but it is possible to study the effects of a naturally occurring earthquake: Oishi et al. (2018) studied the effects of a naturally occurring earthquake essentially by comparing people who completed an online survey just before versus after the Great East Japan Earthquake. Likewise, it would be unethical and unfeasible for researchers to randomly assign people to remain in school for an extra year or leave school a year early, but it is possible to study the effects of laws that increase the minimum school-leaving age: Davies et al. (2018) exploited the raising of the minimum school-leaving age in the United Kingdom in 1972 essentially by comparing people born immediately before September 1957 (i.e., not affected by the reform) with those born in or immediately after September 1957 (i.e., affected by the reform).

Furthermore, demand effects should, on average, affect natural experiments less than they affect randomized experiments. The main difference between a

randomized and natural experiment is that the researcher induces the treatment in a randomized experiment but not in a natural experiment. Thus, the treatment can only induce any hypothesis-related expectations in a randomized experiment but not in a natural experiment. In addition, in randomized experiments, the outcome variable is typically assessed right after treatment administration because researchers are the ones applying the treatment. In natural experiments, researchers do not administer the treatment. Thus, the outcome variable is often measured months or years after the naturally occurring treatment has taken place. For example, cognitive abilities were measured decades after the minimum school-leaving age had been increased in the United Kingdom (Davies et al., 2018). Thus, the researchers' hypotheses are usually less obvious and demand effects are usually less of an issue in natural experiments.

In addition, natural experiments overcome the dependence of randomized experiments on the willingness of the participants to be randomized: People do not have a choice about whether or not to be affected by a natural experiment that involves, for example, an earthquake or compulsory military service. They still need to agree to participate in a study investigating the effects of the natural experiment (unless administrative records are used). However, agreeing to participate may be a lower threshold in a natural than in a randomized experiment because the chances of receiving the desired or undesired treatment do not depend on study participation in a natural experiment.

Finally, natural experiments enable researchers to assess whether an effect that they found in a randomized experiment in the lab is detectable and of a relevant size in the populations, contexts, and conditions of interest. The samples from natural experiments tend to be more representative (e.g., systematic vs. convenience sampling), the treatments are real-world events (e.g., military service) rather than artificial manipulations (e.g., exposure to preselected violent stimuli), and the study conditions and contexts are ecologically exemplary.⁴ Thus, natural experiments can be considered valuable complements to randomized experiments in triangulation efforts (i.e., the application of multiple approaches to causal inference in which each approach has different strengths, weaknesses, and sources and directions of bias; e.g., Hammerton & Munafò, 2021). For example, Cesario (2022) argued that traditional laboratory experiments in social psychology may not inform us about real-world group disparities. To address this issue, social psychologists could complement randomized experiments with studies exploiting natural experiments. For all these reasons, we consider natural experiments to be attractive complements to randomized experiments that might be particularly helpful in evaluating whether causal effects are of relevant size in the field.

Types of Studies Exploiting Natural Experiments

In line with Dunning (2012; see also Sieweke & Santoni, 2020), we distinguish between three types of studies exploiting natural experiments: standard natural experiments, instrumental-variable designs using a natural experiment, and regression-discontinuity designs (see Fig. 1). For the first two of these three types of studies, we distinguish between random assignment to treatment/instrument and as-if random assignment to treatment/instrument. Random assignment means that participants are assigned to the treatment/instrument through a randomization process with a known probability distribution. As-if random assignment means that participants are not assigned through an actual randomization process. However, because of the natural occurrence of the event/condition that constitutes the treatment/instrument, neither the experimenter nor the participant/unit have control over the treatment/instrument in a natural experiment with as-if randomization. For example, self-selection into treatment and control conditions is not possible. Nevertheless, units might be selected by someone/something (e.g., the government) into treatment and control conditions on the basis of factors that are related to the units (e.g., income, age). Thus, when using natural experiments with as-if randomization, it is particularly important to consider potential confounding factors because the absence of random assignment can lead to confounding bias in standard natural experiments and violations of assumptions in instrumental-variable designs using a natural experiment. For the third type of study that exploits natural experiments, regression-discontinuity designs, we distinguish between designs that are sharp and those that are fuzzy. The decision tree in Figure 2 illustrates how to determine whether an event or condition is a suitable natural experiment and which type of design can be used to exploit it.

Standard natural experiments

In an example for a standard natural experiment with random assignment, Lindqvist et al. (2020) compared Swedish lottery players who won a large sum of money with matched controls (i.e., lottery players who had the same sex, age, and number of lottery tickets but who did not win a large sum) to estimate the effect of wealth on well-being (Box 1). Other examples are public-policy interventions designed to be equitable by randomly allocating costs (e.g., Vietnam lottery draft) or benefits (e.g., Green Card Lottery), as well as randomized admission procedures (e.g., for medical school; Ketel et al., 2016). The standard natural experiment with random assignment is most similar to randomized experiments as routinely implemented by psychological researchers.

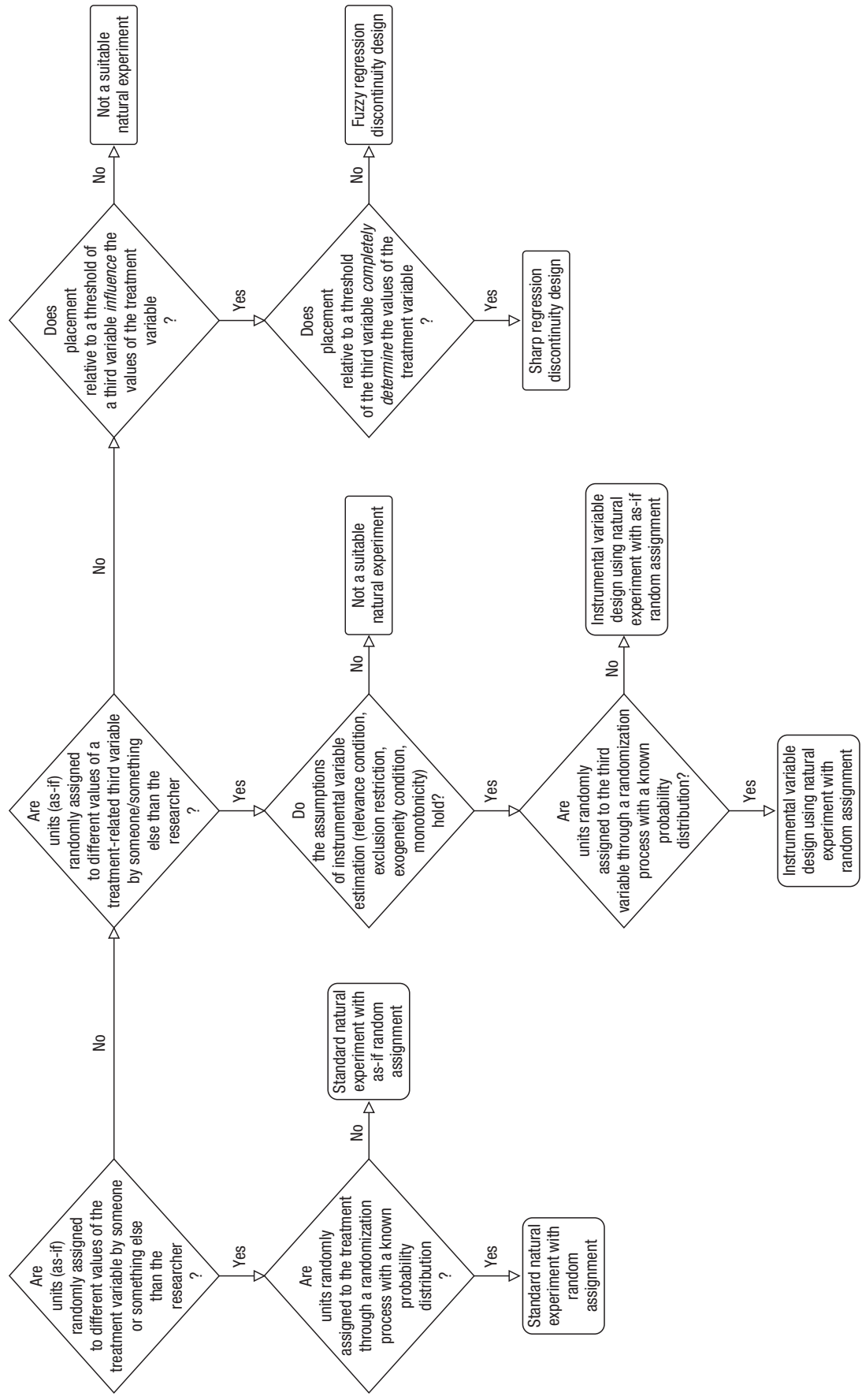


Fig. 2. Decision tree for identifying (types of) natural experiments.

Examples of natural experiments with as-if randomization are naturally occurring events, such as famines or earthquakes (e.g., Belloc et al., 2016); policy interventions, such as the U.S. 1970 Clean Air Act amendments (Schwaba et al., 2021); alphabetical seating orders (e.g., Byrne, 1961); biological sex of a child or sibling (e.g., Dudek et al., 2022); or the randomization of offspring genotype during meiosis (i.e., offspring as-if randomly inherits one allele from each parent at every point in the genome; e.g., Madole & Harden, 2023).⁵

As already alluded to above, in the case of as-if randomization, it is particularly important to check for potential confounding factors that might bias the effect estimation. Dudek et al. (2022) provided an example of how such checks might be performed. In their analysis of data from 85,887 people from 12 diverse samples covering nine countries, Dudek et al. aimed to estimate the causal effect of growing up with a next-younger sister rather than a next-younger brother on one's personality. To be able to estimate this effect, Dudek et al. first needed to establish that the gender of the next younger sibling was as-if random. Although whether a younger sibling will be a male or female is essentially random at conception, sex differences in survival rates or sex-selective abortions could mean that the sex of the next younger sibling will not be completely random. To rule out such artifacts, Dudek et al. performed several balance checks to assess whether the two groups (people with a next younger sister vs. people with a next younger brother) were comparable before the birth of the next younger sibling. For example, they compared the two groups on a number of observable variables (e.g., number of older siblings) that were determined before the birth of the next younger sibling. If the gender of the next younger sibling was indeed random, there should be no systematic differences between the groups on such variables (including unobservable ones). Finally, they performed robustness checks by excluding data from three samples for which there were concerns about sex-selective abortion and other imbalances. Overall, they concluded that the sex of the next younger sibling was plausibly as-if random, and thus, they could use it to estimate the effect of the siblings' sex on personality.

Instrumental-variable design using a natural experiment

A naturally randomized variable can serve not only as a treatment variable, as is the case in standard natural experiments, but also (alternatively) as an instrumental variable. An instrumental variable allows one to target variation in the treatment that is plausibly unconfounded. In that manner, an instrumental variable (Z) allows researchers to unbiasedly estimate the causal effect of some other treatment variable (X) on an outcome (Y),

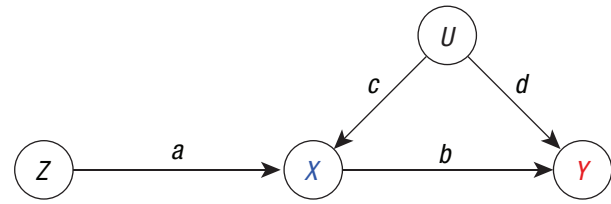


Fig. 3. Directed acyclic graphs illustrating instrumental-variable estimation. The regression estimate β_{YX} is biased because it is influenced not only by the effect of X on Y but also by the effects of U on X and Y : $\beta_{YX} = b + cd$. If certain assumptions hold (see the main text and, e.g., Lousdal, 2018), the instrumental-variable estimate $\frac{\beta_{YZ}}{\beta_{XZ}}$ is unbiased.

By taking the ratio of the strength of the association between Z and Y to the strength of the association between X on Z , the estimate isolates the covariation in X and Y that is causal: $\beta_{XZ} = a$, $\beta_{YZ} = ab$, and $\frac{\beta_{YZ}}{\beta_{XZ}} = \frac{ab}{a} = b$. Z does not necessarily need to have a direct causal effect on X (as is the case here) for the instrumental-variable estimation to work (for details, see Fig. S1 on the OSF at <https://osf.io/a5nmx>). All variables in the above equations need to be standardized except for the ratio of the instrumental-variable estimator. The figure and formulas were inspired by Lousdal (2018) and Thoemmes (2022). Z = instrumental variable; X = treatment variable; U = unobserved confounding variable; Y = outcome variable.

even if there is unobserved confounding between X and Y (U ; see Fig. 3).

A classic example for an instrumental-variable approach was provided by Angrist (1990). He used draft eligibility determined by the Vietnam lottery draft (Z) as an instrument to estimate the causal effect of veteran status (X) on civilian earnings (Y). Such instrumental-variable estimation can work even when veteran status and civilian earnings are confounded by unobserved third variables, such as civilians' earning potential (U)—men with few civilian job opportunities are more likely to enlist. Not only a naturally randomized variable (natural experiment) but also a variable randomized by a researcher (randomized experiment) can be used as an instrumental variable. However, the focus of the current article is exclusively on naturally randomized variables (natural experiments).

When does this approach work? Instrumental-variable estimation produces unbiased estimates of the effect of X on Y under several assumptions (e.g., Bollen, 2012; Lousdal, 2018; Wooldridge, 2010), among them the following three central ones⁶:

- **Relevance condition:** Z has a causal effect on X (Fig. 3). For example, draft eligibility according to the lottery has an effect on veteran status.⁷
- **Exclusion restriction:** Z affects the outcome Y only through X . Draft eligibility has an effect on civilian earnings only via veteran status.
- **Exchangeability or exogeneity condition:** Z does not share unobserved common causes with Y . Draft eligibility and civilian earnings have no unobserved common causes.

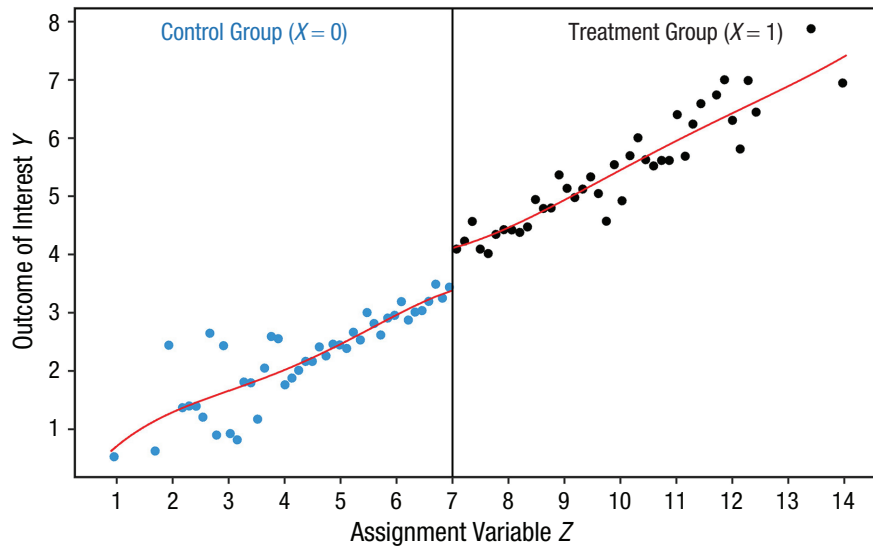


Fig. 4. Illustration of a treatment effect in a regression-discontinuity design. We used the R package `rdrobust` (Version 2.1.0; Calonico et al., 2022) and `ggplot2` (Version 3.3.6; Wickham, 2016) to plot simulated data illustrating a treatment effect on the outcome of interest. The vertical line indicates the threshold of the assignment variable (threshold value = 7). Units to the left of the vertical line (values < 7) did not receive the treatment; units to the right of the line (values > 7) received the treatment. We simulated data for 1,000 units. The graph displays the means of 100 bins, each containing 10 units. The R code for the simulation and plot can be found on the OSF at <https://osf.io/a5nmx>.

The three central assumptions might alternatively be met when Z does not have a direct causal effect on X but when X and Z share a common cause U^* (for details, see Fig. S1 on the OSF at <https://osf.io/a5nmx>).

Only the relevance condition can be empirically tested. Natural experiments can help to meet the exchangeability condition because the random or as-if random assignment ensures that there are no (unobserved) common causes of Z and Y . For example, being eligible for the draft and civilian earnings should have no common causes because the birthdays that were drawn in the lottery were randomly determined (unless the lottery was rigged).

Green and Winik (2010) applied instrumental-variable estimation to determine the effect of incarceration (X) on recidivism (Y ; Box 2). They made use of the natural experiment that drug offenders were as-if randomly assigned to one of nine different judicial calendars (Z) in the U.S. District of Columbia Superior Court in 2002 and 2003. Because each calendar came with a different group of judges, and the judges varied in terms of how punitive/lenient they were, the sentences from the nine calendars varied in terms of prison and probation time, more so than one would expect by chance alone: The least punitive calendar incarcerated 23% of the offenders, and the average prison sentence was 5.1 months; the most punitive calendar incarcerated 65% of the offenders, and the average prison

sentence was 11.9 months. Green and Winik used eight dummy variables that indicated which calendar the offender was assigned to (Z) as instruments for incarceration (X) on recidivism (Y), a crucial piece of information for legal proceedings in which psychologists often provide expert opinions on matters of imprisonment and release.

Regression-discontinuity design

The third type of study using a natural experiment is the regression-discontinuity design. In a regression-discontinuity design, a threshold of an assignment variable Z determines whether or not a unit receives treatment X (Fig. 4). The main identifying assumption in a regression-discontinuity design is that potential-outcome conditional-expectation functions are continuous at the discontinuity (i.e., at the threshold). In other words, the treatment variable changes discontinuously as a function of the assignment variable Z at the threshold, whereas the covariates (including unobserved confounders) change continuously or do not change at all as a function of the same assignment variables Z (i.e., the distributions of the covariates are a smooth function at the threshold; e.g., Hahn et al., 1999). In this case, causal identification is possible despite units not being (as-if) randomly assigned to the treatment variable X .

For example, Thistlethwaite and Campbell (1960) investigated the effect of public recognition (X) on subsequent academic achievement (Y) by comparing students who received certificates of merit with similar students who did not (Box 3). Students received a certificate of merit if their test score on the Scholarship Qualifying Test (Z) was at least equal to the qualifying score in the student's state. Students whose scores equaled the qualifying score were classified by Thistlethwaite and Campbell in Interval 11, and those whose scores were 1 unit lower than the qualifying score were classified in Interval 10. The assumption of continuous confounder distribution was plausible only for students near the qualifying score in the student's state. For example, students in Interval 11 were probably very similar in their pretreatment characteristics to students in Interval 10, even more so if the exam did not have perfect reliability and thus introduced some random variability into the exam scores. Therefore, when comparing students in Intervals 10 to students in Interval 11, any confounding influences are arguably negligible, and marked differences between these groups in the outcome variable (i.e., subsequent academic achievement) could actually be attributed to the treatment (i.e., recognition via certificate of merit).

The assumption of continuity of the confounder distribution is plausible only if units (e.g., students) in a close neighborhood around the threshold have no direct control over the assignment variable. That is, students just below the threshold cannot manipulate their test scores in such a way that they still obtain the certificate of merit. This assumption cannot be tested directly; violations could be suggested by a lack of balance on covariates and a bunching of units on one side of the threshold of the assignment variable (e.g., Cattaneo et al., 2020; McCrary, 2008; Thoenmes et al., 2017).

Typical thresholds in a regression-discontinuity design are population- and size-based thresholds, such as class size in school and number of employees (e.g., antidiscrimination law that applies only to firms with at least 15 employees; Hahn et al., 1999). Other thresholds are time-based thresholds (e.g., if the time at which a survey was completed was before vs. after an earthquake; Oishi et al., 2018), age or birth date (e.g., school-entry age cutoffs; Ritchie & Tucker-Drob, 2018), eligibility criteria (e.g., college admission cutoff scores on high school exit exams; Dasgupta et al., 2022), and indices (i.e., a composite score that combines information from several variables; Dunning, 2012).

The kind of regression-discontinuity design that we have portrayed so far is sometimes called a sharp regression-discontinuity design because the placement relative to the threshold completely determines whether the treatment is received. In a fuzzy regression-discontinuity

design, the placement relative to the threshold influences the receipt of the treatment but does not determine it completely. In such cases, to estimate the causal effect, the assignment variable is used as an instrumental variable, and the assumptions of instrumental-variable estimation need to hold (see above).

For example, Kuehnle and Oberfichtner (2020) used a fuzzy regression-discontinuity design to estimate the effects of universal childcare (X) on children's cognitive skills and personality traits (Y). They capitalized on the fact that the time at which a child was enrolled in childcare in West Germany was influenced by the calendar year in which the child turned 3. Many children born toward the end of a year start receiving childcare in the summer *before* their third birthday; by contrast, many children born at the beginning of the subsequent calendar year start receiving childcare in the summer *after* their third birthday. Thus, the average age at which children began receiving childcare jumped discontinuously by 5 months between the birth months December and January in the data analyzed by importantly, the December/January threshold did not affect the time at which the children began formal schooling, and the children on the two sides of the threshold did not differ in relevant observable characteristics, such as the mother's native language or education. Because the probability of starting childcare earlier did not switch from 0 to 1 between December and January, their approach was a fuzzy rather than a sharp regression-discontinuity design. Kuehnle and Oberfichtner used a dummy as an instrumental variable predicting the age at which children began childcare to indicate whether a child was born before or after the December/January threshold (Z). They used this instrumental variable to estimate the causal effect of starting universal childcare 4 months earlier, around age 3, on the responses to standardized cognitive tests, Big Five personality questionnaires, and other measures at age 15.

As another example, Gauriot and Page (2019) used a fuzzy regression-discontinuity design to estimate the momentum effect, that is, the effect of the success of sports behavior (X) on subsequent sports performance (Y). They exploited the fact that the probability of whether a player will win a point in tennis varies discontinuously as a function of the location of the ball on the court. The player loses a point if the ball hit by the player lands just outside the court lines. Conversely, the play continues if the ball lands just inside the court lines, giving the player a chance to win the point. Gauriot and Page extracted a very small share of points for which the ball bounced within a few centimeters of the court lines from a large data set on precise ball location during tennis matches between professional tennis players. Because the location of the ball does not completely determine whether a player will win or lose a point—the

play continues if the ball lands inside—their design was a fuzzy rather than a sharp regression-discontinuity design. Gauriot and Page predicted winning a point (X) by a dummy instrumental variable indicating whether the ball landed inside or outside the court lines (Z). They used this instrumental variable to estimate the causal effect of winning or losing a point on the player's subsequent performance.

The Challenges of Using Natural Experiments

Despite their advantages, natural experiments are not a panacea. Like many other causal-inference methods, natural experiments rely on assumptions that are often challenging or impossible to validate empirically. Identifying and analyzing natural experiments often requires profound knowledge and an understanding of subjects that fall outside of many psychologists' core areas of expertise. For example, it might be necessary to know the level of compliance with a policy reform and how it was implemented to evaluate whether the reform can serve as a suitable natural experiment that meets the required assumptions (e.g., Lillebø et al., in press). Likewise, it might be necessary to have a deep understanding of biology and genetics (e.g., issues such as pleiotropy and assortative mating) to evaluate whether and how the random allocation of genetic variants from parents to their children can be used as natural experiments to estimate causal effects (e.g., Madole & Harden, 2023; Sanderson et al., 2022).

Another notable challenge is finding a suitable natural experiment for a particular study. Even if a relevant natural experiment can be found, it might not precisely constitute the treatment of interest or it might only allow researchers to identify a particular causal effect. For example, in Dudek et al. (2022), the effect of the sex of the next younger sibling was identified but not the effect of the sex of the next older sibling or of any other sibling. Out of substantive considerations, all of these effects would be relevant to provide a full picture of how siblings shape personality. Thus, focusing exclusively on natural experiments would narrow down the causal effects that could be studied by psychologists.

Nonetheless, we believe that natural experiments offer attractive opportunities for estimating causal effects, especially when randomized experiments are unethical, unfeasible, or generalize poorly to the populations, contexts, and conditions of interest. Thus, we hope that the current work and the examples highlighted herein will inspire readers to be on the lookout for suitable applications in their own work, thus leading to the wider use of natural experiments in psychology and, ultimately, a broader toolbox for causal inference in our field.

Transparency

Action Editor: David A. Sbarra

Editor: David A. Sbarra

Author Contributions

Michael P. Grosz: Conceptualization; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Writing – original draft; Writing – review & editing.

Adam Ayaita: Conceptualization; Investigation; Methodology; Writing – review & editing.

Ruben C. Arslan: Conceptualization; Investigation; Writing – review & editing.

Susanne Buecker: Conceptualization; Investigation; Writing – review & editing.

Tobias Ebert: Conceptualization; Investigation; Writing – review & editing.

Paul Hünermund: Investigation; Methodology; Writing – review & editing.

Sandrine R. Müller: Conceptualization; Writing – review & editing.

Sven Rieger: Conceptualization; Investigation; Writing – review & editing.

Alexandra Zapko-Willmes: Conceptualization; Investigation; Writing – review & editing.

Julia M. Rohrer: Conceptualization; Investigation; Methodology; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding


This work was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project Number 461127198.


Open Practices


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 iDs

Michael P. Grosz  <https://orcid.org/0000-0002-1949-4384>

Susanne Buecker  <https://orcid.org/0000-0003-3443-5400>

Tobias Ebert  <https://orcid.org/0000-0003-0146-2517>

Paul Hünermund  <https://orcid.org/0000-0001-9163-038X>

Sandrine R. Müller  <https://orcid.org/0000-0002-1226-6370>

Julia M. Rohrer  <https://orcid.org/0000-0001-8564-4523>

Acknowledgments

We thank Jane Zagorski for language editing, Felix Thoemmes for stimulating discussions and suggestions, and Jan Lorenz Westermann for help with the literature review (searching for and coding studies, preparing the R code, etc.). S. Buecker is now at the School of Psychology and Psychotherapy, Witten/Herdecke University, and A. Zapko-Willmes is now at the Department of Psychology, University of Siegen.

Notes

1. The term “quasi-experiment” has been used inconsistently in the literature. Some authors use it as a synonym for natural experiments (e.g., Rutter, 2007). Others have pointed out that unlike natural experiments, quasi-experiments do not come with the presumption that units have been randomly assigned to treatment and control conditions (e.g., Dunning, 2012; Shadish et al., 2002; Sieweke & Santoni, 2020). For example, Shadish et al. wrote: “By definition, quasi-experiments lack random assignment. Assignment to conditions is by means of self-selection, by which units choose treatment for themselves, or by means of administrator selection, by which teachers, bureaucrats, legislators, therapists, physicians, or others decide which persons should get which treatment” (p. 14). Nevertheless, natural experiments with as-if random assignment resemble quasi-experiments in that they also lack assignment to treatment and control conditions through a randomization process with a known probability distribution.
2. A directed acyclic graph illustrates the assumptions about the causal connections between a set of variables (e.g., Pearl, 2009; Textor, 2023). For example, an arrow from X to Y indicates that a direct causal effect of X on Y is assumed.
3. For each of the three research areas (psychology vs. business economics vs. sociology), we conducted two electronic database searches on Web of Science on October 13, 2023. In the two searches, we searched the abstracts, titles, and keywords of publications from the years 2000 to 2023 for the following search terms: ALL = (“natural experiment” OR “instrumental variable” OR “regression discontinuity”) AND (“study” OR “studies” OR “experiment” OR “data”) and ALL = (“study” OR “studies” OR “experiment” OR “data”), respectively. In the search for natural experiments, we included not only the search term “natural experiment” but also “instrumental variable” and “regression discontinuity” because the latter two represent two common types of natural experiments (see Fig. 1). A screenshot of the search results can be found on the OSF at <https://osf.io/a5nmx>.
4. Whether natural experiments have higher generalizability to the populations, contexts, and conditions of interest than randomized experiments depends on the specific natural experiment and the randomized experiment it is being compared with. Compared with randomized laboratory experiments (e.g., the effects of priming on decision-making in a dictator game), most natural experiments will have higher generalizability. However, compared with field experiments, natural experiments might have similar or even lower generalizability.
5. We considered the randomization of offspring genotype during meiosis to be as-if random rather than truly random because segregation distortions (e.g., transposons and meiotic drive) might be deviations from complete randomness in the inheritance of potential parental alleles (e.g., Fishman & McIntosh, 2019).
6. The assumptions in the main text are known as three central assumptions of instrumental-variable estimation. These three assumptions can alternatively be formulated as two assumptions in the context of directed acyclic graphs: First, there must be an open path between instrument Z and the treatment X , and second, all paths between Z and the outcome Y must be closed in a modified graph in which all edges out of X are removed (e.g., Brito & Pearl, 2002; Textor, 2023). A further assumption not mentioned in the main text is the monotonicity assumption: Z must

not increase X for some individuals and decrease it for others (e.g., Bollen, 2012; Labrecque & Swanson, 2018; Lousdal, 2018). Furthermore, if individual-level causal effects are heterogeneous (i.e., the effect is not identical across all individuals), then the instrumental-variable estimation identifies the local average treatment effect (LATE) rather than the average treatment effect. The LATE is the average treatment effect in the subset of the population whose treatment selection is induced by the instrument (e.g., Morgan & Winship, 2015). For example, the LATE is the average effect of veteran status on civilian earnings among the people who attained veteran status because they were eligible for the draft according to the lottery (excluding people who would have attained veteran status even without the lottery and people who did not attain veteran status despite being eligible for the draft according to the lottery).

7. Valid instruments need to be strong in the sense that their quantitative effect on the treatment needs to be sufficiently large (given the sample size); otherwise, they can produce inconsistent and instable parameter estimates (e.g., Bound et al., 1995). A commonly used threshold for deciding whether instruments are sufficiently strong is a first-stage F statistic exceeding 10 (Staiger & Stock, 1994; but see also Keane & Neal, 2023).

References

- Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review*, *80*(3), 313–336. <https://www.jstor.org/stable/2006669>
- Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.
- Bastardo, N., Matthews, M. J., Sajons, G. B., Ransom, T., Kelemen, T. K., & Matthews, S. H. (2023). Instrumental variables estimation: Assumptions, pitfalls, and guidelines. *The Leadership Quarterly*, *34*(1), 101673. <https://doi.org/10.1016/j.leaqua.2022.101673>
- Belloc, M., Drago, F., & Galbiati, R. (2016). Earthquakes, religion, and transition to self-government in Italian cities. *Quarterly Journal of Economics*, *131*(4), 1875–1926. <https://doi.org/10.1093/qje/qjw020>
- Bollen, K. A. (2012). Instrumental variables in sociology and the social sciences. *Annual Review of Sociology*, *38*, 37–72. <https://doi.org/10.1146/annurev-soc-081309-150141>
- Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, *90*(430), 443–450. <https://doi.org/10.1080/01621459.1995.10476536>
- Bruto, C., & Pearl, J. (2002). Generalized instrumental variables. In *UAI'02: Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence* (pp. 85–93). Morgan Kaufmann Publishers. <https://dl.acm.org/doi/10.5555/2073876.2073887>
- Byrne, D. (1961). The influence of propinquity and opportunities for interaction on classroom relationships. *Human Relations*, *14*(1), 63–69. <https://doi.org/10.1177/001872676101400106>
- Calonic, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2022). *rdrobust: Robust data-driven statistical inference in regression-discontinuity designs* (Version 2.1.0)

- [Computer software]. CRAN. <https://CRAN.R-project.org/package=rdrubust>
- Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2020). *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press.
- Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2023). A practical introduction to regression discontinuity designs: Extensions. ArXiv. <https://doi.org/10.48550/arXiv.2301.08958>
- Cesario, J. (2022). What can experimental studies of bias tell us about real-world group disparities? *Behavioral and Brain Sciences*, *45*, Article e66. <https://doi.org/10.1017/S0140525X21000017>
- Corneille, O., & Lush, P. (2023). Sixty years after Orne's *American Psychologist* article: A conceptual framework for subjective experiences elicited by demand characteristics. *Personality and Social Psychology Review*, *27*(1), 83–101. <https://doi.org/10.1177/10888683221104368>
- Dasgupta, U., Mani, S., Sharma, S., & Singhal, S. (2022). Effects of peers and rank on cognition, preferences, and personality. *Review of Economics and Statistics*, *104*(3), 587–601. https://doi.org/10.1162/rest_a_00966
- Davies, N. M., Dickson, M., Davey Smith, G., Van Den Berg, G. J., & Windmeijer, F. (2018). The causal effects of education on health outcomes in the UK Biobank. *Nature Human Behaviour*, *2*(2), 117–125. <https://doi.org/10.1038/s41562-017-0279-y>
- Diener, E., Northcott, R., Zyphur, M. J., & West, S. G. (2022). Beyond experiments. *Perspectives on Psychological Science*, *17*(4), 1101–1119. <https://doi.org/10.1177/17456916211037670>
- Dudek, T., Brenøe, A. A., Feld, J., & Rohrer, J. M. (2022). No evidence that siblings' gender affects personality across nine countries. *Psychological Science*, *33*(9), 1574–1587. <https://doi.org/10.1177/09567976221094630>
- Dunning, T. (2012). *Natural experiments in the social sciences: A design-based approach*. Cambridge University Press.
- Ertola Navajas, G., López Villalba, P. A., Rossi, M. A., & Vazquez, A. (2022). The long-term effect of military conscription on personality and beliefs. *Review of Economics and Statistics*, *104*(1), 133–141. https://doi.org/10.1162/rest_a_00930
- Fishman, L., & McIntosh, M. (2019). Standard deviations: The biological bases of transmission ratio distortion. *Annual Review of Genetics*, *53*, 347–372. <https://doi.org/10.1146/annurev-genet-112618-043905>
- Galizzi, M. M., & Navarro-Martinez, D. (2019). On the external validity of social preference games: A systematic lab-field study. *Management Science*, *65*(3), 976–1002. <https://doi.org/10.1287/mnsc.2017.2908>
- Gauriot, R., & Page, L. (2019). Does success breed success? A quasi-experiment on strategic momentum in dynamic contests. *The Economic Journal*, *129*(624), 3107–3136. <https://doi.org/10.1093/ej/uez040>
- Gelman, A., & Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, *37*(3), 447–456. <https://doi.org/10.1080/07350015.2017.1366909>
- Green, D. P., & Winik, D. (2010). Using random judge assignments to estimate the effects of incarceration and probation on recidivism among drug offenders. *Criminology*, *48*(2), 357–387. <https://doi.org/10.1111/j.1745-9125.2010.00189.x>
- 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. <https://doi.org/10.1177/174569162092152>
- Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201–209. <http://www.jstor.org/stable/2692190>
- Hahn, J., Todd, P. E., & van der Klaauw, W. H. (1999, May). *Evaluating the effect of an antidiscrimination law using a regression-discontinuity design* (Working Paper No. 7131). National Bureau Of Economic Research. <https://doi.org/10.3386/w7131>
- Hammerton, G., & Munafò, M. R. (2021). Causal inference with observational data: The need for triangulation of evidence. *Psychological Medicine*, *51*(4), 563–578. <https://doi.org/10.1017/S0033291720005127>
- Heckman, J. J., & Moktan, S. (2020). Publishing and promotion in economics: The tyranny of the top five. *Journal of Economic Literature*, *58*(2), 419–470. <https://doi.org/10.1257/jel.20191574>
- Huntington-Klein, N. (2022). *The effect: An introduction to research design and causality*. CRC Press. <https://doi.org/10.1201/9781003226055>
- Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, *79*(3), 933–959. <https://doi.org/10.1093/restud/rdr043>
- Keane, M., & Neal, T. (2023). Instrument strength in IV estimation and inference: A guide to theory and practice. *Journal of Econometrics*, *235*(2), 1625–1653. <https://doi.org/10.1016/j.jeconom.2022.12.009>
- Ketel, N., Leuven, E., Oosterbeek, H., & van der Klaauw, B. (2016). The returns to medical school: Evidence from admission lotteries. *American Economic Journal: Applied Economics*, *8*(2), 225–254. <https://doi.org/10.1257/app.20140506>
- Kuehnlé, D., & Oberfichtner, M. (2020). Does starting universal childcare earlier influence children's skill development? *Demography*, *57*(1), 61–98. <https://doi.org/10.1007/s13524-019-00836-9>
- Labrecque, J., & Swanson, S. A. (2018). Understanding the assumptions underlying instrumental variable analyses: A brief review of falsification strategies and related tools. *Current Epidemiology Reports*, *5*, 214–220. <https://doi.org/10.1007/s40471-018-0152-1>
- Lillebø, O. S., Markussen, S., Røed, K., & Zhao, Y. (in press). *Not a flying start after all? Journal of Political Economy*.
- Lindqvist, E., Östling, R., & Cesarini, D. (2020). Long-run effects of lottery wealth on psychological well-being. *Review of Economic Studies*, *87*(6), 2703–2726. <https://doi.org/10.1093/restud/rdaa006>
- Lousdal, M. L. (2018). An introduction to instrumental variable assumptions, validation and estimation. *Emerging Themes in Epidemiology*, *15*, Article 1. <https://doi.org/10.1186/s12982-018-0069-7>

- Madole, J., & Harden, K. (2023). Building causal knowledge in behavior genetics. *Behavioral and Brain Sciences*, 46, Article E182. <https://doi.org/10.1017/S0140525X22000681>
- McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2), 698–714. <https://doi.org/10.1016/j.jeconom.2007.05.005>
- Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.
- Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.
- Oishi, S., Kohlbacher, F., & Choi, H. (2018). Does a major earthquake change attitudes and well-being judgments? A natural experiment. *Social Psychological and Personality Science*, 9(3), 364–371. <https://doi.org/10.1177/1948550617707016>
- Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). Cambridge University Press.
- Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science*, 29(8), 1358–1369. <https://doi.org/10.1177/0956797618774253>
- Rutter, M. (2007). Proceeding from observed correlation to causal inference: The use of natural experiments. *Perspectives on Psychological Science*, 2(4), 377–395. <https://doi.org/10.1111/j.1745-6916.2007.00050.x>
- Sanderson, E., Glymour, M. M., Holmes, M. V., Kang, H., Morrison, J., Munafò, M. R., Palmer, T., Schooling, C. M., Wallace, C., Zhao, Q., & Davey Smith, G. (2022). Mendelian randomization. *Nature Reviews Methods Primers*, 2, Article 6. <https://doi.org/10.1038/s43586-021-00092-5>
- Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, 13(4), 279–313. <https://doi.org/10.1037/a0014268>
- Schwaba, T., Bleidorn, W., Hopwood, C. J., Gebauer, J. E., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2021). The impact of childhood lead exposure on adult personality: Evidence from the United States, Europe, and a large-scale natural experiment. *Proceedings of the National Academy of Sciences, USA*, 118(29), Article e2020104118. <https://doi.org/10.1073/pnas.2020104118>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for causal inference*. Houghton Mifflin Company.
- Sieweke, J., & Santoni, S. (2020). Natural experiments in leadership research: An introduction, review, and guidelines. *Leadership Quarterly*, 31(1), Article 101338. <https://doi.org/10.1016/j.leaqua.2019.101338>
- Staiger, D. O., & Stock, J. H. (1994). Instrumental variables regression with weak instruments. *Econometrica*, 65(3), 557–586. <https://doi.org/10.2307/2171753>
- Textor, J. (2023). *Drawing and analyzing causal DAGs with DAGitty: User manual for DAGitty*. <https://dagitty.net/manual-3.x.pdf>
- Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6), 309–331. <https://doi.org/10.1037/h0044319>
- Thoemmes, F. (2022, June 27–28). *Kausale Effekte in der Persönlichkeitspsychologie* [Causal effects in personality psychology] [Paper presentation]. Explicit Causal Inference in Personality Research Network 4th Meeting, Mannheim, Germany.
- Thoemmes, F., Liao, W., & Jin, Z. (2017). The analysis of the regression-discontinuity design in R. *Journal of Educational and Behavioral Statistics*, 42(3), 341–360. <https://doi.org/10.3102/1076998616680587>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (Version 3.3.6) [Computer software]. <https://ggplot2.tidyverse.org>
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). MIT Press.