Longitudinal and Life Course Studies • vol 11 • no 1 • 7–25
 • Paper received April 2019/Revised August 2019/Published online November 2019
Online ISSN 1757-9597 • https://doi.org/10.1332/175795919X15659210629362

RESEARCH ARTICLE

Causality in life course research: the potential use of 'natural experiments' for causal inference

Ross Macmillan[1], *ross.macmillan@ul.ie*
Carmel Hannan, carmel.hannan@ul.ie
University of Limerick, Ireland

Recent decades have seen renewed attention to issues of causal inference in the social sciences, yet implications for life course research have not been spelled out nor is it clear what types of approaches are best suited for theoretical development on life course processes. We begin by evaluating a number of meta-theoretical perspectives, including critical realism, data mining and experimentation, and find them limited in their potential for causal claims in a life course context. From this, we initiate a discussion of the logic and practice of 'natural experiments' for life course research, highlighting issues of how to identify natural experiments, how to use cohort information and variation in the order and timing of life course transitions to isolate variation in exposure, how such events that alter social structures are the key to identification in causal processes of the life course and, finally, of analytic strategies for the extraction of causal conclusions from conventional statistical estimates. Through discussion of both positive and negative examples, we outline the key methodological issues in play and provide a road map of best practices. While we acknowledge that causal claims are not necessary for social explanation, our goal is to explain how causal inference can benefit life course scholarship and outline a set of practices that can complement conventional approaches in the pursuit of causal explanation in life course research.

Key words theory • Research Methods • causality • natural experiments

To cite this article: Macmillan, R. and Hannan, C. (2020) Causality in life course research: The potential use of 'natural experiments' for causal inference, *Longitudinal and Life Course Studies*, vol 11, no 1, 7–25, DOI: 10.1332/175795919X15659210629362

At first blush, one might think that studies using a life course perspective would have particular affinity with questions of causal processes and, by extension, causal inference. Questions of temporality, stability and change, trajectories and transitions, and the foundational idea of lives unfolding in unique socio-historical contexts are all central features of a life course *perspective* and all have parallels in discussions of causality and causal inference (see, for example, Elder et al, 2003; Morgan and Winship, 2015). At the same time, it is not clear how concerned life course *research* is about making credible causal claims or about adopting research strategies that allow for better causal inference given well-recognised and unchallenged threats to causal claims. As there are clearly a number of advantages to research that can credibly make causal claims, life course researchers may be missing important opportunities for extending their scope of research.

Life course researchers may, however, have particular advantages with respect to causal inference. In particular, longitudinal research designs where the same people are followed over time may be better at measuring outcomes given the greater use of prospectively recorded information and better suited to ordering data in time. Hence, such data can satisfy basic dicta that causes precede effects in time. Such data also allow researchers to identify changes in personal experiences that can be associated with changes in personal outcomes. Here, people can serve as their own controls and researchers can also mitigate potential biases associated with the universe of time-stable attributes of individuals (Halaby, 2004). Yet, longitudinal data may still fall victim to issues of selection and endogeneity that introduce bias into metric coefficients and hence undermine causal inference (Heckman, 1981; Angrist and Pischke, 2008). The key issue is systematic differences between the 'treatment' and 'control' groups, those stratified by some explanatory variable, which will bias estimates of the variable on some outcome. So while life course research may have conceptual frameworks and data that facilitate causal inference, typical analytic strategies fall short.

With this background, this paper has three objectives. First, we articulate three different approaches to causality and discuss their relevance for life course research. Second, we discuss the potential contributions of 'natural' experiments for causal inference with specific attention to how they are identified, how longitudinal data can be organised to exploit their value, and how statistical approaches, including difference-in-difference estimation, regression discontinuity and instrumental variables estimation, facilitate causal inference. Finally, we provide some synthesising discussions and make a case for combining the rich archives of longitudinal data with exogenous changes in social structure to produce better ideas of cause and effect in life course processes.

Conceptualising causality: juggling nihilism, pragmatism and dogmatism

Three broad perspectives on causality adopt different normative and operational approaches to causality and causal inference. The first, critical realism, asserts that much of reality exists and operates independently of our awareness or knowledge of it (Danermark et al, 2005). Empirical social science, in contrast, is based on the ability to observe variables and processes directly and to employ a variety of discursive strategies for articulating cause and effect. The problem articulated by Bhaskar (2013) and others (Archer et al, 2013) is that coincidence between an independent and dependent variable may tell us very little about the mechanisms of causal influence. The problems are myriad, but mechanisms can go inactivated (because of the necessity of a third (fourth, fifth and so on) factor as in the case of a statistical interaction), mechanisms can be activated but not be perceived (due to the finite nature of observation), or be activated but be negated by some countervailing force. In the end, critical realism argues that limits of observation and measurement means that the counterfactual necessary to prove causal effects can never be fully established.

Works at the opposite end of the spectrum, those focused solely on the organisation of data have also cast some doubt on the value of causal identification. Andrew Abbott, for example, has argued that current conceptions of causality fall far short of their theoretical objectives (Abbott, 1988; 1998). An alternative strategy (see Abbott and Hrycak, 1990) is to focus more attention on identifying salient and informative patterns in data. Such a strategy has more than a few variants, but appears in the life course literature most prominently in form of latent class analysis (Macmillan and Copher, 2005) and sequence analysis (Aisenbrey and Fasang, 2010; Elzinga and Liefbroer, 2007). The key logic of both approaches is to forgo distinguishing causes and effects and instead to model how variables fall into identifiable configurations that tell theoretical stories.

While philosophers and data miners have retreated from efforts to identify causes, many, if not most, economists have doubled down. Although traditional economics was diverse in its acceptable/preferred empirical methods, recent years have seen the rise of random control trials (RCTs) as its go-to methodology (Deaton, 2009). The key problematic is that any consequential variable that is not included in the model introduces a correlation between the error term and the measured explanatory factors and hence biases coefficients that link the explanatory factors to the outcome (Angrist and Pischke, 2008). With RCTs, the process of randomisation ensures that the only observed or unobserved difference between groups is exposure to whatever treatment (cause) is under study.

While life course scholars interested in questions of causality clearly have options, there seem to be limits in all of the above strategies. In the case of critical realism, the force of argument is really critique and the proposed solutions do not seem to lead anywhere productive. Data mining clearly has value, but efforts to incorporate such work in life course research have left many scholars wanting (see discussions in Levine, 2000; Wu, 2000). A key concern is that sequences and clusters may not contribute much to efforts to *explain* social phenomena. Moreover, variables included in clusters, classes and sequences may in fact have causal relations among them that are simply ignored for empirical convenience. So, data mining neither supplants concerns about causal relationships nor provides a fully palatable alternative.

Although frequently described as the 'gold standard' for social research, RCTs may actually be the worst of the options for life course research. There are four problems. The first is that much and perhaps most of what life course researchers care about is not amenable to experimental manipulation. We cannot randomly assign people to families or relationships and randomly assigning people to other social institutions like work and school or to communities is frighteningly expensive and administratively prohibitive. A second problem is that experiments live or die based on their external validity, yet RCTs almost never have a random sample and instead focus on *in situ* populations where unseen and often long-standing selection dynamics are unknown and perhaps unknowable (Sampson, 2008). Third and related, treatment effects are not homogenous (that is, uniform across the underlying population) and instead are magnified or minimised depending on the non-random attributes of the underlying sample (Heckman, 2001; Manski, 1999; Smith, 2003). This of course moves the ball back into the court of observational studies, with all their problems of causal inference. Finally, life course scholars care about social structure. Yet, the idea of the classic experiment is that social structure is removed (Sampson, 2008). Exposure to a treatment occurs in a social void. Bursztyn and colleagues (2017), for example, conduct a lab-based RCT where the treatment of interest is a world where supervisors and managers in organisations do not know the gender of the person communicating with them. How realistic is this? Or more worrisome, what does it mean to produce a set of findings that occur in a social structure that simply doesn't exist in the real world? These are real world problems that are not solved by naïve advocacy for a method based in its ability to deal with one, and only one, problem of causal inference. In the end, three distinct but influential approaches to questions of causality do not seem to provide much leverage for causally inclined life course researchers.

On the potential of natural experiments for life course research

An alternative with perhaps more promise is the use of 'natural' experiments. In what follows, we describe what natural experiments are and how researchers might go about finding them; illustrate how cohort comparison – a frequent feature of life course data – provides unique opportunities for causal claims; emphasise how attention to the trajectories and transitions provides unique leverage for causal inference; and discuss analytic strategies that, when coupled with the above, provide statistical estimates from which one can more credibly make causal claims.

Identifying natural experiments

The central premise of a natural experiment is the identification of a situation or circumstance that mimics the core dimensions of an RCT. With a true RCT, researchers randomly allocate individuals to a treatment and control group and then manipulate some condition to expose the treatment group to something that the control group does not experience. In theory, the two groups are identical on both observable and unobservable characteristics (pretreatment equality) and differ only on whatever the treatment is or does. In simplest terms, the end result is that any existing differences between the two groups are due to the treatment and only the treatment.

Natural experiments arise when comparable individuals or groups are sorted into treatment and control groups by something that is typically beyond their control. They differ from RCTs in that the exposure and allocation is not controlled by the researcher but instead occurs naturally. 'Naturally', in this context, indicates a large variety of processes that occur at a number of levels. For example, one could think of the comparison of life expectancy in North and South Korea as an example of the effects of political systems on health. To do so requires that one agrees that: (1) North and South Koreans were similar to one another at the point of political differentiation; (2) North and South Koreans did not choose their political systems; and (3) that there is not social, economic and cultural mixing that would contaminate the treatment.

The latter assumption is probably easiest given the strict border controls, censorship of media and hypersegregation of the two populations. The first assumption might be more challenging but has at least has some degree of historical credibility. The most challenging is the middle assumption given the historical legacy of the Korean conflict and its origins in ideological differences around multiple dimensions of social organisation. Still, the example lays out the logic and provides the terms of evaluation.

In general, natural experiments incorporate some aspect of historical change, variation in social structure, or socio-economic or socio-political anomaly into the research framework. In doing so, they often focus on some quasi-random event in the world as a vehicle for telling causally credibly stories of social experience and human development. This yields a first principle in the use of natural experiments for life course research:

(1) A key assumption is that the groups under study are for comparable and for all intents and purposes identical (or conditionally identical once controlling for sorting variables) and that one group is randomly affected by a treatment that is outside their control.

Although we cannot claim comprehensive coverage and in fact are always amazed at the creativity that goes into the use of natural experiments, we identify four types of natural experiments. A first is randomisation in public administration. Here, governments often use randomisation to either ensure equity in the provision or expectations of services. In Finkelstein and colleagues (2012) for example, a natural experiment occurred when the state of Oregon opened a waiting list for Medicaid and then provided insurance to a random subset of applicants. In a similar vein, Angrist (1990) took advantage of the fact that the US government ran televised draft lotteries during the Vietnam War based on day of birth to study the effects of military service on labour market outcomes. Although we delve into the study in more detail later, a more recent example of this approach is Gangl and Ziefle's (2015) study of the impact of changes in family leave and benefits policies on women's commitment to work and family.

Second, natural experiments are often exactly that: variation in exposure because of variation in nature. Kirk (2009), for example, uses the devastation wrought by Hurricane Katrina on the city of New Orleans, USA, and the parishes surrounding it to study the impact of residential relocation on likelihood of recidivism among people recently released from penal institutions. The key feature of the research is the fact that those released in the year prior to the hurricane could choose to move back to their original neighbourhood or to move to some other neighbourhood while those released post-hurricane really could not move back to their original neighbourhood as the neighbourhoods had been destroyed.

Third, social dislocations often produce remarkable and dramatic changes in life circumstances. As one example, Card (1990) studied the impact of a massive increase in low-wage labour on local economies by exploiting the 1980 Mariel Boatlift. In spring 1980, the Castro government opened the port at Mariel to anyone who wished to leave, with the consequence that more than 125,000 Cubans arrived in the Miami area over a period of a few months. For Card, the key issues for causal inference lie in (1) the unplanned and unexpected nature of the influx, (2) the stark increase in the size of the Miami labour force, approximately 7%, and (3) the fact that such increases were not seen in other surrounding cities. Through comparison with these cities, Card could speak to the question of whether wages and employment rates are suppressed by the influx of low-wage, immigrant labour (and they are not).

Finally, heterogeneity in the availability of ICT and other technology often produces large differences in the availability of information with important social consequences. Kearney and Levine (2015), for example, studied the effect of the television show *Sesame Street* on children's test scores. At the time, federal licensing laws dictated that the very popular show was randomly accessible depending on availability of either VHF (very high frequency) or UHF (ultra high frequency). Capitalising on this variation, the researchers were able to compare academic outcomes across places with different levels of reception (and found that those in areas of greater exposure were less likely to get left behind in school).

Exploiting variation by cohort and transitions

If the first dimension of natural experiments is identifying some exposure of significance, the second dimension involves identification of appropriate comparison groups. Akin to arguments for RCTs, the key assumption or realisation is that the exposure of interest is differentially distributed across groups that are, for all intents and purposes, identical. Here, a life course perspective provides guidance. One of the central principles of life course research is the idea that lives unfold in distinct historical contexts (Elder et al, 2003).

Methodologically, researchers activate this principle by drawing on concepts like generation and cohort to differentiate the timing of people's experiences to changing historical conditions (see, for example, Shanahan et al, 1997). Here, different cohorts have distinct experiences because they 'encounter' history at unique points in their lives and this gives them unique experiences. Thus one can simply stratify some group by year of birth, identify the key experience of interest and model some outcome in relation to the intersection of the two.

Elder's (2018) classic work on children of the Great Depression exemplifies this approach and provides a useful template for research on a host of topics. Importantly, the empirical underpinnings of this approach provide a route for research with the potential for much stronger claims for causal inference.

The key starting point is cohort differences in life experience. There is no question that cohort differences exist and can be hugely consequential. At the same time, the social structures around cohorts are often remarkably stable. Most of the major institutions in society and the rules and resources that characterise them do not change easily or quickly (Sewell, 1992). As a consequence, cohorts that are adjacent to one another typically inhabit worlds that are very similar. From the standpoint of empirical research, one can argue that adjacent cohorts are, for all intents and purposes, identical to one another. Obviously, this is always an empirical question.

But life course scholarship is organised on the principle that social differentiation occurs due to difference in exposures. If no differences exist or occur, cohort differentiation should not exist. This leads us to a second principle for using natural experiments for causal inference:

> (2a) Cohort comparison can help identify causal effects if one can credibly claim or show that such cohorts are/were indistinguishable prior to or in the absence of exposure to a cause of interest.

If cohorts present one potential avenue of comparison, exploitation of life course dynamics presents another. Life course scholarship places particularly weight on the unfolding of human lives and how the life course can be conceptualised as patterned transitions through social institutions (Shanahan, 2000). Traditionally, this involves transition into and out of formal schooling; movement into paid employment; entry into intimate partnerships, be they marriages or cohabitation; parenthood; and transitions out of one's childhood home. The value of this in the use of natural experiments is that it yields interesting and often multiple comparison groups for both 'controls' and the 'treated'. Returning to Kirk (2009), the key issue of interest is knowing whether high rates of recidivism reflect continuity of community and social networks after one is released from prison. The problem is disentangling selection into neighbourhoods from neighbourhood effects in and of themselves. To do this, Kirk focuses on people who had been convicted of crimes and incarcerated while residing in the metropolitan area around New Orleans. The key life course dynamic that allows him to create treatment and control groups is the *timing* of release and whether it occurred before or after Hurricane Katrina levelled large swathes of New Orleans and the surrounding area. He uses timing to construct two 'control' groups that were released prior to the hurricane and one 'treatment' group released post-hurricane. The logic of the analysis is that the post-hurricane group will have had their relocation choices highly restricted, would be significantly less likely to return to their original community in New Orleans, and hence will have access to pre-incarceration criminal associates restricted or constrained.

Another example that capitalises on variation in life course dynamics to specify treatment and control groups is a recent study by Gangl and Ziefle (2015). Here, the question of interest is whether changes in parental leave entitlements in Germany altered the balance of psychological commitments to work and family. In 1992, the duration of benefits was extended from 18 months to 36 months with benefit entitlements extended to the 24th, month and this stayed in place with minimal change through 2007. Using the German Socioeconomic Panel (GSOEP), the researchers were able to identify changes over time associated with exposure to different policy environments, specifically prior to (1989–1991) or after (1992–2000 and 2001 and beyond) the policy change. What facilitates causal claims is identification of multiple 'treatment' and 'control' groups. On the treatment side, mother's exposure to the policy is shaped by whether they were employed or not at the time the policy was enacted. Mothers who were employed were entitled to both the parental leave and a flat rate benefit; mothers who were not employed were only entitled to the benefit (that is, there was no job to take a leave from). On the control side, one comparison group is young women without children, while a second is older women with completed fertility. The key assumption is not that the groups are substitutable in work–family orientations (as there is no reason to assume this) but that their trends over time would be similar were it not for variation in the policy environment (see section on 'Difference-in-difference estimation', later). In general, such work highlights a complementary principle:

> (2b) The order and timing of key life course transitions generates important and multiple treatment and control groups that can facilitate both assessment of the robustness of results and potentially illuminate mechanisms linking cause and effect.

What is a cause? Social structural opportunities and measurement

The identification of 'treatment' and 'control' groups requires a clear understanding of the 'exposure' of interest. Yet, here we think that much of the literature has been overly cavalier and yields a number of cautionary tales. Although there are numerous examples, we focus on a highly influential study of education and health: Lleras-Muney (2005). Her set-up was extremely elegant. She identified discrete years in which laws were passed in certain US states that mandated longer participation. She then used this to identify 'treatment' groups where options for leaving school were legally constrained and 'control' groups where they were not. With this set-up, Lleras-Muney could extract two very valuable pieces of information. First, do cohorts indeed have different educational attainment depending on whether they were exposed to the participation laws? Second, do cohorts have different mortality based on whether they were exposed to the participation laws? If the answer to these questions is yes, the two differences can be calibrated to quantify just how much mortality decreases given changes in educational attainment. In Lleras-Muney's research design, she focused attention on synthetic cohorts who were 14 years old between 1914 and 1939. Tracking these cohorts over time, one can calculate death rates during specific intervals in the future, for example between 1960 and 1970 and between 1970 and 1980 and then examine whether death rates correlate with both exposure to compulsory attendance laws or law induced-educational attainment. With this set-up, Lleras-Muney reports large, statistically significant effects and interprets the coefficients as causal effects of education on mortality risk.

To us, Lleras-Muney is a cautionary tale about the need to think deeply about what constitutes a cause. Any experiment, natural or other, lives or dies based on the quality of the treatment. In this case, the treatment is exposure to compulsory attendance laws and its knock on for average educational attainment. So it is useful to look at the laws to see what exactly they are doing. In examining each law in Lleras-Muney's (2005) appendix B, one learns some interesting things. First, there are a not-trivial number of cases where compulsory attendance was relaxed and these cases are not used. Second, many laws were relatively short lived. California, for example, switched back and forth between six and seven mandated years of schooling five times between 1921 and 1936.[2] Maybe this is not a problem, but there certainly appears to be SUTVA (stable unit treatment value assumption) issue given that people in the very same school and even in the very same family are likely treated quite differently under the ever-changing laws. Third, and perhaps most striking, little education was actually mandated. The overwhelming majority of laws mandated seven years of mandatory schooling with almost all of the laws tightly clustered between six and eight. The mandated change is also quite small. In the vast majority of cases, the laws increased compulsory attendance by one or two years. In other words, these laws were mandating either completion of primary school or some secondary school attendance (with secondary schools becoming core parts of the education system between 1910 and 1930). If one believed nothing more than the human capital model of education (Becker, 1994) or its application to a 'health production function' (Grossman, 1999), perhaps an additional year of education at the end of primary / beginning of secondary school would matter. But for most social scientists, this seems like a particularly trivial part of the attainment distribution and the overall conclusion that each additional year of compulsory education increases educational attainment by 4.8% is not particularly compelling.

The problem is one of social structure. Educational attainment may be adequately measured as years of formal education but the *meaning* of educational attainment requires attention to the operation of educational and other institutions, institutions which are themselves embedded in unique social and historical contexts. Not being experts on US educational policy and practice in the early 20th century, we simply do not know what changes in cognition, behaviour and social circumstance accrues from adding one year of education that results in six to seven years of total schooling. There are obviously dozens of potential connections between greater education and health in general (see, for example, Ross and Wu, 1995), but it is unclear whether any of them would be activated by this amount of change in educational attainment. Given this, it may not be surprising to learn that a broad, multinational assessment of 18 cases of compulsory schooling reform for 12 European countries between 1903 and 1976 found next to no support for a causal effect of any magnitude (Gathmann et al, 2015). Out of a possible 98 effects, only 14 for males and only 5 for females were statistically significant. If one were to correct for multiple assessments using a Bonferroni adjustment, it is not clear that any would be statistically significant. Such cautionary tales indicate a third principle in the use of natural experiments for life course research:

> (3) Causes and resulting 'treatments' should be conceptualised at the level of social structures and specifically capture social dynamics that meaningfully alter the conditions of everyday life in ways that are theoretically salient for some given outcome.

Although it may seem like a trite principle, it is useful to note that the dominant trend in the use of natural experiments is to focus on some change in policy that is largely decontextualised, assumed to have homogeneous meaning and effects, and is often transported from one study to another given its proven ability of satisfying the core requirements of instrumental variables. The use of compulsory education laws to identify causal effects of educational attainment for example has been used to study a host of health measures and risk behaviours (see review in Eide and Showalter, 2011), wages, earnings and wealth (Devereux and Hart, 2010; Oreopoulos, 2006), marital stability (Oreopoulos and Salvanes, 2011), fertility (Black et al, 2008; Cygan-Rehm and Maeder, 2013), criminal offending and incarceration (Bell et al, 2016), political interest, involvement and voting (Milligan et al, 2004), life satisfaction and happiness (Oreopoulos, 2007), religious affiliation, church attendance and frequency of praying (Hungerman, 2014), and belief in superstition and horoscopes (Mocan and Pogorelova, 2017). It is also worth noting that the only areas where there have been sustained attempts to replicate are the areas of wages and health and at best the weight of the evidence is quite mixed on both counts. To us, this is expected given that education was quite heterogeneous in structure and process at the time of most reforms (although this rapidly changed in the 1980s and 1990s with the consequence that educational systems are remarkably similar in the contemporary period (Meyer et al, 2017)) and that the consequences of education continue to be loosely coupled with coarse measures of attainment.

Drawing inference

While the identification of a 'natural experiment' opens the door for causal estimation, problems still exist in that treatment and controls remain dynamic in the face of a natural experiment. Here, we discuss three strategies for drawing causal inference with natural experiments. These include difference-in-difference approaches, regression discontinuity and instrumental variable analysis.

Difference-in-difference estimation

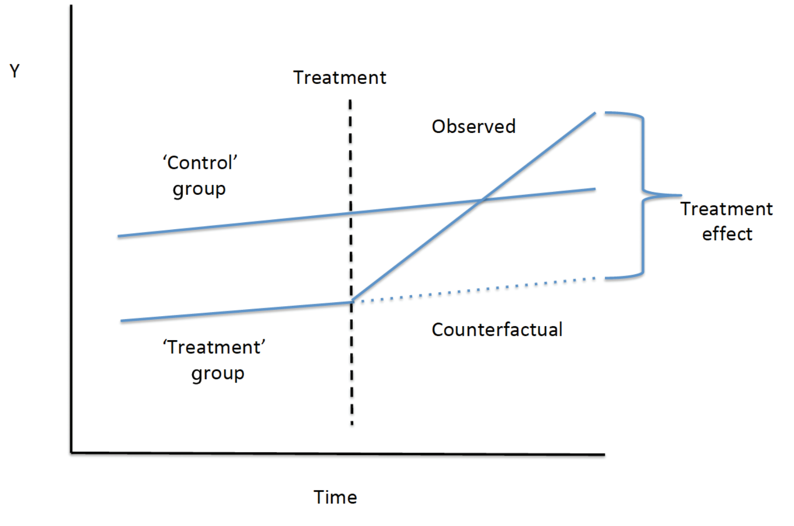
Difference-in-difference (hereafter D-I-D) estimation is arguably the simplest strategy, but is also the one most vulnerable to threats to causal inference. The logic of D-I-D is that one tracks changes over time in an outcome as a way of identifying a causal effect. The key identifying assumption is that the exposure introduced through the natural experiment occurs in one population and not another. Given this, a causal estimate is identified through change occurring in a treatment population *relative* to change occurring in the control group. For causal inference, the two populations must have pretreatment equality and there are a number of strategies for assessing this. Pretreatment equality does not imply that the two groups are identical on the outcome. Indeed, groups can be widely disparate on the outcome, but they need to have parallel trends prior to the treatment. Figure 1 provides a graphic description of this with y-axis measuring the value of the outcome of interest, Y, and the x-axis capturing time.

Here, the treatment and control groups have different levels prior to the treatment but parallel trends over time. The magnitude of this difference *is stable over time*. With longitudinal data and more than one data point, one can formally assess this either by looking at differences at successive points in time and assessing degree of change or, alternatively, by adopting a regression approach and seeing whether a coefficient capturing change over time varies across groups. In the post-treatment period, the key assumption is that the treatment group has a counterfactual that is the expected value on the outcome were there no change in trend among the treatment group. There is no way to validate this assumption, but its strength is directly proportional to the strength of the parallel trend assumption and by extension pretreatment equality.

Determination of the treatment effect is simply the difference between the counterfactual value and observed value. Depending on the type of data one has, there are two complementary strategies for deriving a causal estimate. One is simple algebra. One can calculate the difference in outcome value before and after for the treatment and the control groups and then subtract the differences from one another. There is an analogue using standard regression techniques. Formally,

$$y = \beta_0 + \beta_1 T + \beta_2 S + \beta_3 (T * S) + \varepsilon$$

Figure 1: Simple difference-in-difference set-up



where y is the outcome, T is a dummy variable indexing period (pre-exposure = 0, post-exposure = 1), S is a dummy variable indexing group (not exposed = 0, exposed = 1), and T∗S is a composite capturing the unique value on the outcome for the exposed group in the post-exposure period. With a regression approach, one can include further covariates that might confound the relationship between exposure and outcome or one can elaborate the model by differentiating within the two groups to further sharpen the group of interest (for example, a 'triple difference' model). Again, Gangl and Ziefle (2015) are useful in that their analyses explicitly differentiate work–family orientation for women with and without completed fertility before and after changes in family leave and benefit policy in Germany.

As a final comment, there is a technical literature on statistical issues around D-I-D approaches but most of this focuses on issues of efficiency rather than bias. For us, the critical issue is the parallel trend assumption. With either strong argumentation or strong evidence, one can reasonably assume that the difference between the exposed and non-exposed groups in the pre-exposure period would have remained unchanged were it not for the exposure. In the absence of such argument or evidence, D-I-D estimates have reasonable credibility as causal effects.

Regression discontinuity

A second approach for improving causal inference is regression discontinuity (hereafter RD). With regression discontinuity, there exists some threshold that crisply separates groups into an exposed group and a non-exposed group. In the classic work of Thistlethwaite and Campbell (1960), the focus of attention was the impact of scholarships on school performance. The central problem is that scholarships are awarded based on merit and those who received them are high performers prior to the award and should be high performers after the award. The solution was to focus only on students on either side of the award line. In other words, if students were awarded a scholarship if their grade point averages (GPA) exceeded 80%, then attention would focus solely on those with GPAs between 79% and 81% with the assumption that such groups would be very similar to one another and comparison of school performance for those with and without scholarships within these groups would yield the causal effect of scholarship awards. In life course research, there are frequently age cut-offs for a variety of dynamics and researchers could exploit such cut-offs for causal inference. Similarly, adjacent cohorts can be seen as functionally similar to one another while simultaneously subject to different exposures.

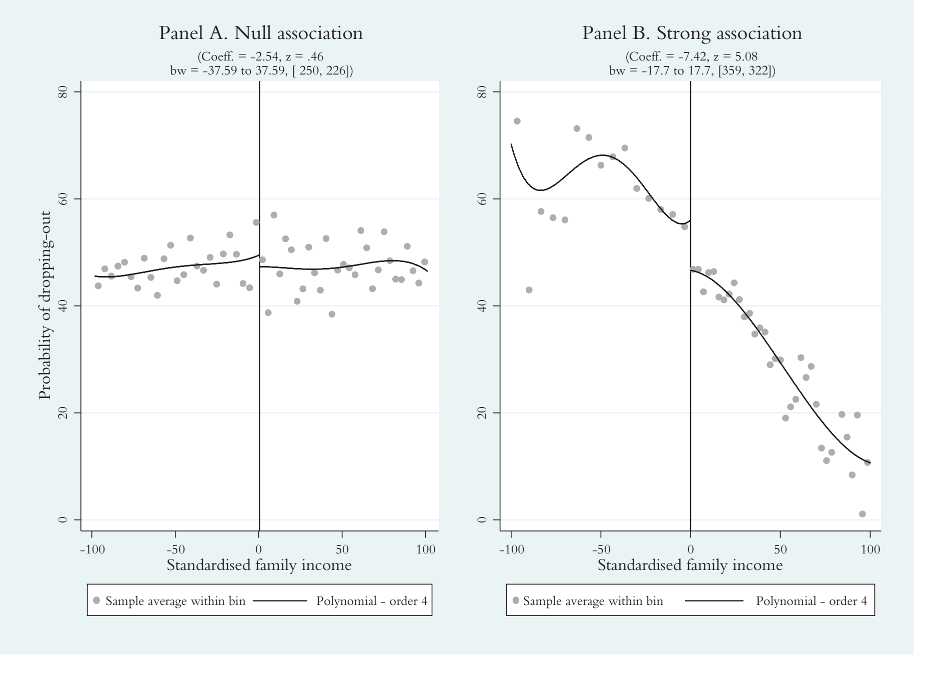
RD relies on three principles. First, the research subjects must not be able to manipulate their status through recognition of the threshold. For example, a student who is able to convince a teacher to award them a higher score that moves them across the threshold may differ in some observable way from those who cannot. This would introduce selection bias into the process.

Second, there needs to be some degree of random error in the threshold metric. If grading was perfect and involved no error, there would be systematic differences between those that have GPAs of 79% and those that have GPAs of 81%. But if one assumes some degree of random error in grading, as is generally accepted, then one can assume that the 79%ers and the 81%ers are really the same population and hence the scholarship is independent of pre-existing traits.

Third, there presumably is some 'bandwidth' where cases within the 'band' are sufficiently similar to one another and where cases outside the band are sufficiently different. Returning to the GPA example, it may be perfectly acceptable to assume that those within the 79% to 81% band are substitutable for one another, yet highly problematic to assume something similar for those between 70% and 90%. In practice, establishing an appropriated bandwidth is crucial and there is important work that provides guidance (Imbens and Kalyanaraman, 2012). Given the importance of these assumptions, RD diagnostics typically focus on examining cases on either side of the threshold as to their similarity or difference.[3]

Figure 2 shows examples of RD analyses using the *rdrobust*, *rdbwselect* and *rdplot* in Stata 15.1. The example is adapted from Calonico et al, (2014). Here, the running variable is a standardised measure of parental income, the threshold is the income cut-off for eligibility of financial support, and the outcome is the likelihood of dropping out of college. Panel A is the association between two random variables with a uniform distribution. The key aspect of the graphic is the lack of a gap at the threshold (0) and the apparent similarity of values, average values and variances on either side of the threshold. This results in a non-significant coefficient of −2.54 (z = .46) for the bandwidth of −37.59 to +37.59 on the parental income measure. For an original sample, the analysis uses 250 cases below the threshold and 226 cases above the threshold. Panel B shows the case where there is a strong discontinuity at the threshold. This is the more realistic example where there is a strong gradient to parental income and the likelihood of attaining a college degree. Yet from the perspective of RD, most of this information is irrelevant to the estimation of the causal effect due to unobserved heterogeneity bias. Instead,

Figure 2: Regression discontinuity example



attention focuses on the cases on either side of the threshold. In this case, there is a clear gap with cases immediately below the threshold having significantly lower average values than cases immediately above the threshold. Again, we can formally estimate this difference by establishing an appropriate bandwidth, restricting analyses to cases within the band, and calculating the difference in average values for cases on either side of the threshold. In this case, the resulting coefficient has a statistically significant value of −7.46 (z = 5.08) for cases within the bandwidth of −17.7 to 17.7. Of the 1,297 cases in the original sample, the RD analysis is restricted to 359 cases below the threshold and 322 cases above the threshold.

RD designs are not common in life course research. One recent example, however, is a paper on the impact of politics and religion on fertility by Aksoy and Billari (2018). The natural experiment in play is the electoral success of the pro-family, pro-natalist AK Parti in the 2004 Turkish election. In this research, the discontinuity is defined in terms of narrow margins of victory in district elections that is calculated as the size of the vote share for the AK Parti in relation to the next largest party. The logic of the argument is that close elections can be tipped either way by idiosyncratic occurrences (such as bad weather) and hence elections that are particularly close are substantively similar on either side of the threshold. Drawing on the work of Imbens and Kalyanaraman (2012), they adopt a bandwidth of .15 (.075 margin of victory/loss) and conclude that exposure to AK Parti governance increases fertility between 2006 and 2010 by 7.75 children per 1,000 women between the ages of 15 and 49, an increase that is robust to the inclusion of several possible confounders and is specified due to social welfare provisions that targeted young families, mothers and the poor. Other RD examples within a life course tradition include Ludwig and Miller's (2007) study of the impact of Head Start, a multifaceted education-based enrichment programme on life course outcomes in the US and Bernardi's (2014) study of compensatory advantages in grade promotion among primary school children in France.

Instrumental variable analysis

A third approach is a variant on the broad use of instrumental variables (hereafter IVs) for causal identification (Bollen, 2012). Consider, for example, prisoners released in Kirk's (2009) analyses. For those released prior to Hurricane Katrina, three quarters returned to their original parish and one quarter took up residence in a different parish. If the 'treatment' is residence in a different neighbourhood, then fully one quarter selected into the treatment. Even in the post-Katrina period, allocation was not crisp. Here, 50% returned to the same parish and 50% went to a different parish. The latter is actually quite remarkable given the level of devastation produced by the hurricane.

Given issues of selection and non-compliance, it can be useful to treat a natural experiment as an IV for the treatment of interest. With an IV, estimation involves two stages. In the first, a variable that is unrelated to the outcome of interest is used as a predictor (instrument) of the treatment of interest. In the case of Kirk (2009), the period of release pre- or post-Katrina predicts whether people returned to their original parish or not. In the second stage, the outcome of interest is regressed on the predicted treatment measure. This, under ideal circumstances, removes the correlation between the explanatory variable and unobserved characteristics. In doing so, estimation only uses the portion of the variability in the treatment effect that is uncorrelated with omitted variables. The key challenge, one that characterises all IV estimation, is the need to satisfy the 'exclusion' assumption, the assumption that the instrument and the outcome are unrelated to one another (see discussion in Angrist and Pischke, 2014). It is also important to recognise that IV estimation limits generalisability. As Angrist and Krueger (2001: 77) note, IVs provide an estimate for a specific group – namely, people whose behaviour can be manipulated by the instrument. In econometric terms, the latter is described as the *local average treatment effect* or LATE.

The obvious challenge with such approaches is satisfying the necessary assumptions. We see three dimensions to this. First, one needs to have a theoretical rationale for the assumptions. This is often the most challenging. Take for example the use of Hurricane Katrina. If the hurricane altered labour market opportunities in general, it could increase the likelihood of rearrest. If the hurricane reconfigured opportunities for crime (that is, it produced more empty residences; Cohen and Felson, 1979), it could have a direct effect on rearrest. If the hurricane mobilised law enforcement as emergency personnel and hence increased surveillance, it could increase likelihood of rearrest. Unfortunately, there is no end to such 'what if' questions and ultimately instruments have some invisible and potentially variable threshold as to their credibility. Second, when possible one should assess the validity of the exclusion restriction. In models that are just identified, when the number of instruments equals the number of endogenous variables, the assumption is not testable. Yet, if the model is overidentified, when there are more instruments than endogenous variables, one can use a Sargan–Hansen test that evaluates the joint null hypothesis that the instruments are valid with the test statistic distributed as a chi-square variable. Third, the instrumental variable must have a reasonably strong association with the endogenous variable and hence it is important to test for weak identification. While some argue that the F-statistics at the first stage should be greater than 10, Stock and Yogo (2005) offer a *bias* and *size* approach that takes into account the number of endogenous regressors, number of instruments, maximum tolerable bias and the estimation procedure.

Conclusions

This paper has been motivated by two general questions. First, given existing and influential arguments about how causality could/should work in the social sciences, how should life course researchers proceed? To answer this question, we have had to ask two other questions. Should we abandon interest in causal relationships among observable factors? Or do we want to base our understanding of the social world on information from artificial environments or contrived and highly problematic versions of the real world? To these latter questions, our answers are unequivocal nos. We lose a lot by abandoning inquiry into causal relationships and RCTs, in the lab or in the field, are not a panacea. Second, given this situation, how might we exploit changes in the socio-historic conditions of everyday life to draw better causal conclusions about social dynamics over the life course? In answering this question, we focus attention on the potential of 'natural experiments' and illuminate a small set of best practices.

Given this, our goals in this paper are modest. The goal of true, generalisable causal inference may be difficult, perhaps impossible, to attain, but problem of systematic bias in regression-type analyses is a simple reality. And, the literature provides three responses. First, some retreat from the traditional ideas of causality and instead offer alternative perspectives (critical realism or data mining, for example). Second, some double down and advocate, often aggressively, for approaches that focus on model integrity and internal validity (see, for example, Banerjee and Duflo, 2011). Third, and preferable for no other reasons than feasibility and transparency, is the use of 'natural' experiments where naturally occurring phenomena create situations where reasonably comparable groups have different exposures to something of life course importance. As the latter approach offers something that has theoretical affinity with the guiding principles of life course research, we have described in some detail what natural experiments often look like, how life course data facilitates identification of 'treatment' and 'control' groups, and statistical approaches that facilitate credible identification of effects.

Natural experiments are also not a panacea and instead should be viewed as a complement to, not a replacement of, existing approaches. Some of the more informative research compares effects across models with different assumptions as a way of assessing the credibility of both simple models and more complex assumptions (such as Duke and Macmillan, 2016). At the same time, work such as Kirk (2009) is clearly an important addition in an area of study with a long and detailed history. What is important in such work, however, is that researchers are clear about their assumptions, their scope conditions and the limitations of their work. In the realm of natural experiments, researchers should be particularly careful to avoid reifying particular coefficients as better representations of reality. Estimates of D-I-D are highly dependent on what is being compared with what. RD estimates will vary, often widely, depending on bandwidth. IV estimates live or die based on the quality of the instrument. All of these are choices that a researcher makes and have no a priori answers.

In the end, there are three key takeaways. Life course researchers are clearly oriented to questions of causal influence. They organise their questions around cause and effect.

Descriptions of findings emphasise impact and change. Indeed, causality may be particularly ingrained in life course thinking given its dual pinions of lives shaped by socio-historical context and variability in how lives unfold over time. At the same time, most life course research does not do much to either enhance capacity for causal claims or mitigate the well-recognised sources of bias in the go-to methods of generalised linear models. Yet, this issue in life course research appears more ambivalence than apathy and the key question is what life course scholars could be doing that they currently are not doing to enhance the insights, influence and impacts of our work. Here, the exploitation of natural experiments seems a logical and natural avenue of inquiry. Cohorts are often ideal comparators. Life course transitions often sort people into 'treatments' and 'controls'. Longitudinal studies give us leverage on changes over time in response to changing environmental exposures. Concerns with social structure have given us insight into the universe of causes *and* make us more comfortable with heterogeneity in outcomes that seriously undermine the credibility of those with narrow and dogmatic visions of human agency and who view causal inference as an issue of internal validity. With this toolkit, we are uniquely suited to capitalise on natural experiments and it is to our detriment if we do not.

Notes

[1] Corresponding author: Ross Macmillan, Department of Sociology, University of Limerick, Limerick, Ireland; ross.macmillan@ul.ie. An earlier version of this paper

was presented at the 2018 Annual Meeting of the Society for Longitudinal and Life Course Research, Milan, Italy.

[2] Similarly, Montana switched mandatory years of schooling between six and eight five times between 1921 and 1936. Oklahoma did likewise between 1928 and 1936. Texas switched from zero to four, from four to seven, from seven to four, from four to seven, and from seven to eight over the study period. Florida was even more schizophrenic. It switched 11 mandatory years of schooling from zero to six then six to zero and then zero to seven in a mere five years between 1916 and 1921.

[3] One important diagnostic is McCrary's (2008) density test that examines whether subjects are randomly distributed around the threshold. A second diagnostic is simple covariate comparison. A third diagnostic is a falsification test which examines whether an outcome that is highly unlikely to be influenced by the treatment is in fact influenced. Each approach essentially compares the distribution of cases around the threshold and assesses whether it appears random or not.

Conflict of interest

The authors declare that there is no conflict of interest.

References

Abbott, A. (1988) Transcending general linear reality, *Sociological Theory*, 6(2): 169–86. doi: 10.2307/202114

Abbott, A. (1998) The causal devolution, *Sociological Methods & Research*, 27(2): 148–81. doi: 10.1177/0049124198027002002

Abbott, A. and Hrycak, A. (1990) Measuring resemblance in sequence data: an optimal matching analysis of musicians' careers, *American Journal of Sociology*, 96(1): 144–85. doi: 10.1086/229495

Aisenbrey, S. and Fasang, A.E. (2010) New life for old ideas: the "second wave" of sequence analysis bringing the "course" back into the life course, *Sociological Methods & Research*, 38(3): 420–62. doi: 10.1177/0049124109357532

Aksoy, O. and Billari, F.C. (2018) Political Islam, marriage, and fertility: evidence from a natural experiment, *American Journal of Sociology*, 123(5): 1296–340. doi: 10.1086/696193

Angrist, J.D. (1990) Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records, *The American Economic Review*, 80(3): 313–36.

Angrist, J.D. and Krueger, A.B. (2001) Instrumental variables and the search for identification: from supply and demand to natural experiments, *Journal of Economic Perspectives*, 15(4): 69–85. doi: 10.1257/jep.15.4.69

Angrist, J.D. and Pischke, J.-S. (2014) *Mostly harmless econometrics: An empiricist's companion*, Princeton, NJ: Princeton University Press.

Archer, M., Bhaskar, R., Collier, A., Lawson, T. and Norrie, A. (2013) *Critical realism: Essential readings*, Abingdon: Routledge.

Banerjee, A.V. and Duflo, E. (2011) *Poor economics: A radical rethinking of the way to fight global poverty*, New York: Public Affairs.

Becker, G.S. (1994) Human capital revisited, *Human capital: A theoretical and empirical analysis with special reference to education* (3rd edn), Chicago: The University of Chicago press, pp 15–28.

Bell, B., Costa, R. and Machin, S. (2016) Crime, compulsory schooling laws and education, *Economics of Education Review*, 54: 214–26. doi: 10.1016/j.econedurev.2015.09.007

Bernardi, F. (2014) Compensatory advantage as a mechanism of educational inequality: a regression discontinuity based on month of birth, *Sociology of Education*, 87(2): 74–88. doi: 10.1177/0038040714524258

Bhaskar, R. (2013) *A realist theory of science*, Abingdon: Routledge.

Black, S.E., Dvereux, P.J. and Salvanes, G. (2008) Staying in the classroom and out of the maternity ward? The effect of compulsory school laws on teenage births, *The Economic Journal*, 118(530): 1025–57. doi: 10.1111/j.1468-0297.2008.02159.x

Bollen, K.A. (2012) Instrumental variables in sociology and the social sciences, *Annual Review of Sociology*, 38: 37–72. doi: 10.1146/annurev-soc-081309-150141

Bursztyn, L., Thomas F. and Pallais, A. (2017) "Acting wife": marriage market incentives and labor market investments, *American Economic Review*, 107(11): 3288–319. doi: 10.1257/aer.20170029

Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014) Robust data-driven inference in the regression-discontinuity design, *The Stata Journal*, 14(4): 909–46.

Card, D. (1990) The impact of the Mariel boatlift on the Miami labor market, *ILR Review*, 43(2): 245–57. doi: 10.1177/001979399004300205

Cohen, L.E. and Felson, M. (1979) Social change and crime rate trends: a routine activity approach, *American Sociological Review*, 44(4): 588–608. doi: 10.2307/2094589

Cygan-Rehm, K. and Maeder, M. (2013) The effect of education on fertility: evidence from a compulsory schooling reform, *Labour Economics*, 25: 35–48. doi: 10.1016/j.labeco.2013.04.015

Danermark, B., Ekström, M., Jakobsen, L. and Karlsson, J.C. (2005) *Explaining society: An introduction to critical realism in the social sciences*, Abingdon: Routledge.

Deaton, A.S. (2009) 'Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development', NBER Working Paper No. 14690, National Bureau of Economic Research.

Devereux, P.J. and Hart, R.A. (2010) Forced to be rich? Returns to compulsory schooling in Britain, *The Economic Journal*, 120(549): 1345–64. doi: 10.1111/j.1468-0297.2010.02365.x

Duke, N. and Macmillan, R. (2016) Schooling, skills, and self-rated health: a test of conventional wisdom on the relationship between educational attainment and health, *Sociology of Education*, 89(3): 171–206. doi: 10.1177/0038040716653168

Eide, E.R. and Showalter, M.H. (2011) Estimating the relation between health and education: what do we know and what do we need to know?, *Economics of Education Review*, 30(5): 778–91. doi: 10.1016/j.econedurev.2011.03.009

Elder, G.H. Jr (2018) *Children of the great depression*, New York: Routledge.

Elder, G.H. Jr, Johnson, M.K. and Crosnoe, R. (2003) The emergence and development of life course theory, In J.T. Mortimer and M.J. Shanahan (eds), *Handbook of the life course*, New York: Kluwer Academic/Plenum, pp. 3–19.

Elzinga, C.H., and Liefbroer, A.C. (2007) De-standardization of family-life trajectories of young adults: a cross-national comparison using sequence analysis, *European Journal of Population/Revue européenne de Démographie*, 23(3/4): 225–50.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J.P., Allen, H., and Baicker, K. (2012) The Oregon health insurance experiment: evidence from the first year, *The Quarterly Journal of Economics*, 127(3): 1057–106. doi: 10.1093/qje/qjs020

Gangl, M. and Ziefle, A. (2015) The making of a good woman: extended parental leave entitlements and mothers' work commitment in Germany, *American Journal of Sociology*, 121(2): 511–63. doi: 10.1086/682419

Gathmann, C., Jürges, H. and Reinhold, S. (2015) Compulsory schooling reforms, education and mortality in twentieth century Europe, *Social Science & Medicine*, 127: 74–82.

Grossman, M. (1999) The human capital model of the demand for health, NBER Working Paper No. 7078, National Bureau of Economic Research.

Halaby, C.N. (2004) Panel models in sociological research: theory into practice, *Annual Review Sociology*, 30: 507–44. doi: 10.1146/annurev.soc.30.012703.110629

Heckman, J.J. (1981) Heterogeneity and state dependence, In S. Rosen (ed), *Studies in labor markets*, Chicago: University of Chicago Press, pp. 91–140.

Heckman, J.J. (2001) Micro data, heterogeneity, and the evaluation of public policy: nobel lecture, *Journal of Political Economy*, 109(4): 673–748. doi: 10.1086/322086

Hungerman, D.M. (2014) The effect of education on religion: evidence from compulsory schooling laws, *Journal of Economic Behavior & Organization*, 104: 52–63. doi: 10.1016/j.jebo.2013.09.004

Imbens, G. and Kalyanaraman, K. (2012) Optimal bandwidth choice for the regression discontinuity estimator, *The Review of Economic Studies*, 79(3): 933–59. doi: 10.1093/restud/rdr043

Kearney, M.S. and Levine, P.B. (2015) Early childhood education by MOOC: lessons from Sesame Street, NBER Working Paper No. 21229, National Bureau of Economic Research.

Kirk, D.S. (2009) A natural experiment on residential change and recidivism: Lessons from Hurricane Katrina, *American Sociological Review*, 74(3): 484–505. doi: 10.1177/000312240907400308

Levine, J.H. (2000) But what have you done for us lately? Commentary on Abbott and Tsay, *Sociological Methods & Research*, 29(1): 34–40. doi: 10.1177/0049124100029001002

Lleras-Muney, A. (2005) The relationship between education and adult mortality in the United States, *The Review of Economic Studies*, 72(1): 189–221. doi: 10.1111/0034-6527.00329

Ludwig, J. and Miller, D.L. (2007) Does Head Start improve children's life chances? Evidence from a regression discontinuity design, *The Quarterly Journal of Economics*, 122(1): 159–208. doi: 10.1162/qjec.122.1.159

Macmillan, R. and Copher, R. (2005) Families in the life course: interdependency of roles, role configurations, and pathways, *Journal of Marriage and Family*, 67(4): 858–79. doi: 10.1111/j.1741-3737.2005.00180.x

Manski, C.F. (1999) *Identification problems in the social sciences*, Cambridge, MA: Harvard University Press.

McCrary, J. (2008) Manipulation of the running variable in the regression discontinuity design: a density test, *Journal of Econometrics*, 142(2): 698–714. doi: 10.1016/j.jeconom.2007.05.005

Meyer, J.W., Kamens, D. and Benavot, A. (2017) *School knowledge for the masses: World models and national primary curricular categories in the twentieth century*, New York: Routledge.

Milligan, K., Moretti, E. and Oreopoulos, P. (2004) Does education improve citizenship? Evidence from the United States and the United Kingdom, *Journal of Public Economics*, 88(9/10): 1667–95. doi: 10.1016/j.jpubeco.2003.10.005

Mocan, N. and Pogorelova, L. (2017) Compulsory schooling laws and formation of beliefs: education, religion and superstition, *Journal of Economic Behavior & Organization*, 142: 509–39. doi: 10.1016/j.jebo.2017.07.005

Morgan, S.L. and Winship, C. (2015) *Counterfactuals and causal inference: Methods and principles for social research*, Cambridge: Cambridge University Press.

Oreopoulos, P. (2006) Estimating average and local average treatment effects of education when compulsory schooling laws really matter, *American Economic Review*, 96(1): 152–75. doi: 10.1257/000282806776157641

Oreopoulos, P. (2007) Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling, *Journal of Public Economics*, 91(11/12): 2213–29. doi: 10.1016/j.jpubeco.2007.02.002

Oreopoulos, P. and Salvanes, K.G. (2011) Priceless: the nonpecuniary benefits of schooling, *Journal of Economic Perspectives*, 25(1): 159–84. doi: 10.1257/jep.25.1.159

Ross, C.E. and Wu, C.-L. (1995) The links between education and health, *American Sociological Review*, 60(5): 719–45. doi: 10.2307/2096319

Sampson. R.F. (2008) Moving to inequality: neighborhood effects and experiments meet social structure, *American Journals of Sociology*, 114(1): 189–231.

Sewell, W.H. Jr (1992) A theory of structure: duality, agency, and transformation, *American Journal of Sociology*, 98(1): 1–29. doi: 10.1086/229967

Shanahan, M.J. (2000) Pathways to adulthood in changing societies: variability and mechanisms in life course perspective, *Annual Review of Sociology*, 26(1): 667–92. doi: 10.1146/annurev.soc.26.1.667

Shanahan, M.J., Elder, G.H. Jr and Miech, R.A. (1997) History and agency in men's lives: pathways to achievement in cohort perspective, *Sociology of Education*, 70(1): 54–67. doi: 10.2307/2673192

Smith, H.L. (2003) Some thoughts on causation as it relates to demography and population studies, *Population and Development Review*, 29(3): 459–69. doi: 10.1111/j.1728-4457.2003.00459.x

Stock, J.H. and Yogo, M. (2005) Testing for weak instruments in linear IV regression, In D.W.K. Andrews and J.H. Stock (ed), *Identification and inference for econometric models: Essays in Honor of Thomas Rothenberg*, New York: Cambridge University Press, pp 80–108.

Thistlethwaite, D.L. and Campbell, D.T. (1960) Regression-discontinuity analysis: an alternative to the ex post facto experiment, *Journal of Educational Psychology*, 51(6): 309–17. doi: 10.1037/h0044319

Wu, L.L. (2000) Some comments on "sequence analysis and optimal matching methods in sociology: Review and prospect", *Sociological Methods & Research*, 29(1): 41–64. doi: 10.1177/0049124100029001003