Evidence of Causation—The Contribution of Life Course Research, Part I: Dominant Models of Causal Inference and Their Limitations in Life Course Research

HANS-PETER BLOSSFELD

Abstract

Life course research has been increasingly criticized for relying only on observational data where processes by which subjects select themselves (or are selected) into the states of a causal variable are not under the control of the researcher. The primary objectives of this essay, *the first in a two-part set*, are to discuss two dominant models of causal inference and to identify the uses and limitations of randomized control trials (RCTs) and quasi-experimental designs for answering life course questions.

INTRODUCTION

Life course research has flourished during the past three decades and has diffused into many subfields of the social sciences. In social stratification research, for example, the life course approach has even become the dominant theoretical and methodological orientation (Mayer, 2009, 2015a, 2015b). This paradigm has been so successful because it explicitly respects individuals as agents (or decision-makers), focuses on the unfolding lives in their changing contexts, considers the interdependencies of lives over time (i.e., the ties between parents and their children, careers of partners, interactions among peers, etc.), focuses on long-term effects of critical (mostly early) life experiences, takes into consideration that life course processes are often cumulative (Matthew effects) and that effects of events might not only be constant but often time dependent, and pays attention to events of processes at different aggregation levels (the individual level such as different life course domains; the intermediate level such as family and

Emerging Trends in the Social and Behavioral Sciences.
Robert Scott and Marlis Buchmann (General Editors) with Stephen Kosslyn (Consulting Editor).
 ISBN 978-1-118-90077-2.

household structure, schools, school classes, or organizational structures of companies; and the macro level such as business cycle or the modernization process) (Elder & Shanahan, 2006; Mayer, 1997, 2009; Shanahan, Mortimer, & Kirkpatrick Johnson, 2016). Based on empirical evidences, the life course approach has substantially shaped the development of longitudinal concepts, the specification of mechanisms of change, and the formulation of dynamic models explaining how life course processes work (Buchmann, 1989; Elder & Giele, 2009; Elder, Kirkpatrick Johnson, & Crosnoe, 2004; Mayer, 2009, 2015a).

Despite the impressive successes, life course research—like much other empirical work in the social sciences—has been increasingly criticized for relying only on observational data where the processes by which subjects select themselves (or are selected) into the states of a causal variable are not under the control of the researcher. Estimated effects of causal variables on life course outcomes are therefore subject to systematic influences of other variables (which might create selection bias and confounding), so that the estimated parameters in the analytical models might depart from the true causal effects. Although there are suggestions to use quasi-experimental designs for reducing bias in observational studies, which will be described in the following paragraph, there is always the possibility that observed or unobserved variables associated with the outcome in life course studies have been omitted. The general objection with regard to the estimated parameters from life course models is therefore that they are not very credible and that life course researchers use only a loose "effect" terminology (Holland, 1988). In other words, even in the case when life course researchers have a well-specified theoretical model and think in causal terms about how and why causes produce effects in the life course (Marini & Singer, 1988), their estimated regression parameters cannot be given a causal effect interpretation, at least not without further assumptions (Holland, 1988).

This basic critique of observational (life course) studies comes from the statistical approach to causality popularized by Rubin (1974), Holland (1986, 1988) and many other statisticians as well as sociological methodologists (Sobel, 1995, 2005). Certainly, most people would agree that associational parameters based on observational data cannot simply be equated with causal parameters as they are defined by randomized experiments. However, I challenge the general claim that associational parameters, even if they are estimated on the basis of a well-grounded theoretical process model and high-quality longitudinal data, should have per se a minor scientific value. Rather I think that life course studies, and their replications in different contexts, can provide valuable insights into life course mechanisms, as

demonstrated by the impressive achievements of life course research described earlier.

The primary objective of my essay is to discuss several main models of causal inference. Then I will focus on the uses and limitations of randomized controlled trials (RCT) and quasi-experimental designs for answering life course research questions.

TWO DOMINANT MODELS OF CAUSAL INFERENCE

The goal to find scientifically based evidence for causal relationships leads to design questions, such as which inference model is appropriate to specify the relationship between cause and effect and which statistical procedures can be used to determine the strength of that relationship (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). Two different models of causal inference have dominated the work of practitioners in the social sciences over the past decades: (i) the model of "causation as robust dependence" and (ii) the model of "causation as consequential manipulation."

Causation as Robust Dependence

The "causation as robust dependence" approach—which in multiple regression or path analysis is known as the *control variable* approach (Blalock, 1970; Duncan, 1966) and in the econometric analysis of time-series data as "Granger causation" (Granger, 1969; Johnston, 1972)—starts from the presumption that correlation does not necessarily imply causation, but causation must in some way or the other imply correlation. In this view, the key problem of causal inference is to determine whether an observed correlation of variable X with variable Y, where X is temporally prior to Y, can be established as a "genuine causal relationship." X is called a *genuine* cause of Y in so far as the dependence of Y on X cannot be eliminated through additional variables being introduced into the statistical analysis. Thus, in this approach, causation is established essentially through the elimination of spurious (or non-causal) influences.

Although this approach has dominated the social sciences for several decades, many social scientists consider it as a too limited approach today. First, they think that causal inference should not be limited entirely to a matter of statistical predictability but should include predictability in accordance with theory (Goldthorpe, 2001, p. 3). Theory to some degree might protect against data mining by searching through covariates with the aim to increase estimated precision (Deaton & Cartwright, 2016). Second, since scientists rarely know all of the causes of observed effects or how they

relate to one another, it is not possible to be sure that all other important variables have in fact been controlled for (Shadish, Cook, & Campbell, 2002). A variable X can therefore never be regarded as having causal significance for Y in anything more than a provisional sense.

Based on cross-sectional data, regression models, path analysis (e.g., in the tradition of Blau, Duncan, & Tyree, 1967), or structural equations models (Bollen, 1989) have been widely used in economics and sociology to disentangle cause-and-effect relationships and estimate often quite complex recursive causal models. Although these models are based on strong statistical and analytical assumptions, only few have paid much attention to these assumptions (Freedman, 1992). Since these models are also typically grounded in weak theories (Sørensen, 2009), the control variable approach has not created much understanding of the phenomena under study (Freedman, 1992). These cross-sectional models simply did not appropriately represent the substantive process that generates the data (Lieberson, 1985).

Causation as Consequential Manipulation

The second approach of "causation as consequential manipulation" seems to have emerged as a reaction to the limitations of "causation as robust dependence." Instead of "establishing the causes of effects," Holland (1986, 1988) and Rubin (1974, 1978, 1980) are concerned with "establishing the effects of causes." They make clear that it is more to the point to take causes simply as given, and then to concentrate on the question of how their effects can be securely measured. According to this approach, causes can only be those factors that could serve as treatments or interventions in well-designed experiments or quasi-experiments. Thus, given appropriate experimental controls, if a causal factor X is manipulated, then a systematic effect is produced on the response variable Y. The particular strength of this design is that " … while statements in the form 'Y is a cause of X' are likely to be proved wrong as knowledge advances, statements in the form 'Y is an effect of X,' once they have been experimentally verified, do not subsequently become false." (Goldthorpe, 2001, p. 5).

Understood in this way, causation is always relative in the sense that the specific treatment of X_{tr} and its observed outcome Y_{tr} are compared with what would have happened to the same unit if it had not been exposed to this treatment (counterfactual account of causality). Since it is not possible in the same experiment for a unit to be both exposed and not exposed to the treatment, the conception of "causation as consequential manipulation" leads to what Holland (1986) has called the *fundamental problem of causal inference*. For example, a student who completes a

mathematics program cannot go back in time and complete a different one so that the two outcomes could compared. Thus, the question arises of how researchers can make sure that they get convincing measurements for something that is in fact impossible to measure, that is, the outcome of Y_{con}, if the unit had not been exposed to the treatment (X_{con}) in the same experiment?

In the hard sciences, such as physics or chemistry, it is often relatively easy to conduct strictly controlled laboratory experiments and to demonstrate, based on the qualities of the objects under study (e.g., physical entities), what would have happened (Y_{con}) to the same unit (u) if it had not been exposed to the treatment (X_{con}). In other words, it is often plausible to assume that these objects have a constant response over time (temporal stability) and that the effect of the first treatment is transient and does not affect the study object's response to the second treatment (causal transience). Or one can at least assume that the physical entities or chemical substances respond very similar under certain conditions. In these cases, the causal effect for each study object u, CauEff_u, is then easily defined as $\text{CauEff}_u = Y_{tr} - Y_{con}$. In fact, this model of "causation as consequential manipulation," which is used in well-designed controlled laboratory experiments, has been quite successful in the hard sciences.

In other disciplines such as biology, medicine, or psychology, it is, however, often not possible to assume temporal stability and causal transience at the level of each unit and it is normally impossible to eliminate the impact of confounding influences at the unit level. For these sciences, Rubin and Holland suggested a statistical approach to the fundamental problem of causal inference: rather than focusing on specific study units, this approach estimates an average causal effect [ACE, or an average treatment effect (ATE)] for a population of units: $\text{CauEff} = E(Y_{tr} \mid X_{tr}) - E(Y_{con} \mid X_{con})$, where $E(Y_{tr} \mid X_{tr})$ is the expected value for participants in the treatment group, and $E(Y_{con} \mid X_{con})$ is the expected value for participants in the control group. For this solution to work, however, participants in the treatment and control groups should differ only in terms of treatment group assignment, not on any other variables that might potentially affect their responses. The approach to make sure that this is indeed the case is the randomized experiment, where participants are randomly assigned to the treatment and control conditions, so that one can expect that treatment group assignment would, on average, and over repeated experiments, be independent of any measurement or unmeasured pretreatment characteristics (Cox, 1958; Fisher, 1935). In randomized experiments, treatment assignment and unit response are therefore statistically independent of each other and any kind of bias due to selection or confounding is eliminated.

RANDOMIZED CONTROLLED TRIALS AND THEIR LIMITATIONS

Experiments have a long history in the social sciences (Jackson & Cox, 2013 for an overview), particularly in the subdisciplines of psychology, social psychology, education, and economics. Famous examples are the Hawthorne Studies in the 1920s and 1930s (in organizational research), the Perry Preschool Study in 1962–1967 (in educational research), or the Seattle-Denver Income Maintenance Experiment (SIME/DIME) in the 1970s (in social policy research). It is worth noting that many of the old and influential experiments did not apply randomization.

After a period in the 1980s, during which experimental designs have received less attention, there has been a renewed interest among social scientists, behavioral economists in particular, to use RCT. In a completely randomized experiment, the researcher first draws a random sample of units from the population of interest and then randomly assigns the study units (e.g., individuals, pupils, teachers, and schools) to treatment and control groups, so that both groups are statistically similar with regard to pretreatment characteristics (Jackson & Cox, 2013). After the exposure of study units to treatment and control conditions, the difference of averages in an outcome variable is computed for both groups and then attributed to differences in the treatments rather than any other influences (Schneider *et al*., 2007). Under ideal conditions, the ATE is an unbiased estimator of the causal effect, which also has a standard error (Deaton & Cartwright, 2016). Thus, the likelihood that the observed ATE is due to random error can be assessed by a standard statistical test. The RCT is widely perceived as a simple and effective research tool (Holland, 1986), which allows to discover "what works" in terms of policy interventions (Deaton & Cartwright, 2016). Some people consider the randomized experimental design therefore as the "gold standard" for causal inference (Holland, 1988).

The basic experimental model is remarkable because it makes only few assumptions (Deaton & Cartwright, 2016). No substantive theory or expert knowledge on causal mechanisms and causal structures is required (Holland, 1988). Only a causal variable (a treatment) and an outcome variable have to be specified. No assumptions are necessary about other covariates. The treatment effect can be heterogeneous, that is, the study units can respond differently with regard to the treatment. There are also no assumptions necessary with regard to statistical distributions. Only the existence of means in the treatment and control groups are required. In sum, RCTs are very straight forward methods in order to determine parts of the causal structure, to get convincing effect knowledge and to find out what works (Holland, 1988).

Of course, there are many practical research issues that might undermine the credibility of a causal effect estimated with an RCT. I just mention only a few of these problems here: (i) When the study sample of the RCT is not based on a random sample from the population of interest, the results of the experiment are also not representative of the target population. For example, in many applications, the population of interest cannot be precisely enumerated or accessed (Schneider *et al*., 2007). In these cases, the estimation of an ATE applies only to the study sample and at the time when the trial was executed (Deaton & Cartwright, 2016). This issue is particularly important when experimenters use convenience samples (e.g., if they use university students as study units), so that the RCT outcome might differ from the effect of an experiment that would be based on a random sample from the larger population (Agresti & Franklin, 2007, 170 pp). In these cases, the use of an ATE outside the study group requires further justification. Thus, purposive selection into the study population undermines inference of RCTs in just the same way as does selection in observational studies (Deaton & Cartwright, 2016). (ii) In addition, RCTs assume that the experiment does not alter the behavior of participants (Heckman & Smith, 1995). This goal might be difficult to achieve in social science applications because people often have their own stake in a specific outcome. (iii) If the sample of the study group is small, it might also be difficult to ensure that randomization provides a complete equivalence on all pretreatment characteristics (the so-called balance). However, the balance can be improved by increasing the sample size and by stratified randomization (Deaton & Cartwright, 2016). (iv) In addition, it is also well known that people may not accept their assignment to the treatment or control groups, respond to other participants of the RCT, or may affect the outcome effect, if they are not blinded by specific procedures. (v) In addition, computation and comparisons of means might sometimes be problematic, when there are outliers and the distributions of the groups are skewed (Deaton & Cartwright, 2016), as is typically the case in life course processes (Blossfeld, Golsch, & Rohwer, 2007). (vi) Only under the condition that everyone experiences the same gain (or loss) from the treatment, additional parameters of interest such as the median impact or the fraction of the individuals with a positive impact of the treatment can be estimated (Heckman & Smith, 1995). (vii) As in all social science studies, measurement problems can also undermine the results of an RCT. For example, if imperfect indicators or inappropriate proxy variables are used. (viii) Finally, long-term RCT are confronted with the same attrition problems as long-term observational panel studies. In sum, there are many technical problems connected with an RCT, but most of these issues can be mitigated by adjustments of the RCT design.

In the following, I confine my attention to four major limitations of the "effect of causes design," which are particularly relevant for answering life course research questions. I try to show that some features of RCTs have indeed unappealing implications for life course research.

Lack of Theoretical Structure

Holland (1988) emphasizes that theory is not important in the "effects of causes" framework. He considers theory only as a tentative summary of our current knowledge that is subject to change (Holland, 1988, p. 450). He therefore states that an important feature of an RCT is the opportunity to establish "the effect of A" without further theoretical assumptions. "Theories may come and go, but old, replicable experiments never die; they are just reinterpreted" (Holland, 1988, p. 450). In other words, Holland stresses that causal effects in RCTs are particularly credible because they do not rest on (uncertain) theory. Some of the critics of the "effects of causes" approach call RCTs therefore "theory-free learning machines" (Deaton & Cartwright, 2016). This assumption is crucial and is not easily adopted by many life course researchers (Sørensen, 2009).

The purely operational approach to causal inference has consequences for causal inference because we never look at data directly. We always look at them through our theoretical concepts (Fox, 1992). There is no theory-free observation. Thus, even if theoretical concepts are not made explicit, they still affect the estimated causal effects by unstated assumptions. I would like to make this point clearer with two examples. First, in a randomized experiment, it is necessary to specify at least the treatment and the outcome variables. In education research, for example, when the researcher wants to study student's achievement, many different indicators might be used as an outcome variable in the experiment: competence test scores (for literacy, numeracy, etc.), school marks, years of schooling, educational attainment level, educational transitions to upper secondary school, and so on. The "measured causal effect" of an experiment will therefore vary with the choice of these latent and manifest variables of school achievement. An explicit theoretical specification is therefore necessary or is done at least implicitly. Second, if we do not have a theoretical understanding of the causal relationship between treatment and outcome, the measured causal effect might be of limited use for scientific purposes and policy interventions. For example, Nye, Konstantopoulos, and Hedges (2004) observe in a randomized experiment that there are teacher effects on student achievement. However, they acknowledge in their publications that their design cannot identify the specific characteristics that are responsible for teacher effectiveness. In other words, a lack of a well-specified theoretical model becomes seriously disabling when we try to

interpret and use the experimental results. Heckman and Smith (1995, p. 108) have therefore concluded that " … a research program based on experiments is just a list of programs that 'work' and 'don't work' but no understanding of why they succeed or fail." Thus, theory and previous findings from (observational) studies might not only be important for the identification of what exactly should be studied in an experiment but may also help us to explain why we measure a causal effect in an experiment—even if our theoretical understanding is always provisional. Thus, excellent experiments (as other excellent scientific work) require theory and serious expert knowledge of the subject matter. Life course researchers therefore should not accept the theory-free approach of effect of causes as a "gold standard", but they should begin with well thought-through and clearly specified process models (see the following discussion), derived from life course theory, where the unobservables that underlie the selection and research problem are made explicit (Heckman, 2005, p. 138).

Need to Consider the Context of Experiments

A well-designed RCT maximizes internal validity, that is, it provides an unbiased ATE and "gets the causality right." However, one cannot simply assume that a causal effect established in this way is invariant across different contexts (external validity). Rather our evidences from observational life course studies suggest that in modern societies causal effects differ not only markedly across populations and societal settings but also change over time (e.g., across historical periods, birth cohorts, or the life course). In particular, macro changes (e.g., unemployment rates, growth rates, and changes in the structure of populations) might be relevant for certain causal effects but are hard or even impossible to integrate in an (long-term) experimental design. Thus, there is a need of a time-dependent context knowledge that can only be produced by long-term observational studies or long-term RCT. In other words, causal effects established through experiments (at least) implicitly build on highly specialized economic, cultural, or social structures that enable them (Deaton & Cartwright, 2016, p. 33). These contexts should be made at least explicit. Thus, there is a need to specify in detail whether the causal effect of an experiment has only a local applicability (Deaton & Cartwright, 2016) or whether it also holds (or is credible) outside its original experimental setting. This brings me to the issues of extrapolation (whether the same causal effect will hold elsewhere) and generalization (whether the causal effect holds universally or at least widely) (Deaton & Cartwright, 2016). Again, without any further theoretical understanding and expert knowledge (e.g., from earlier observational studies), even multiple replications of experiments cannot provide a guarantee for the conclusion that the

next experiment in a different setting will work in the same way. Thus, in a response to Holland's statement cited earlier, one could say that replicable experiments might—and indeed often do—die in different contexts. "We can often learn much from coming to understand why replication failed and use that knowledge to make appropriate use of the original findings, not by expecting replication, but by looking for how the factors that caused the original result might be expected to operate differently in different settings"(Deaton & Cartwright, 2016, p. 31).

A good example for the generalization issue is the Perry Preschool experiment in Ypsilanti, USA, which started in the 1960s. In this RCT, 120 Afro-American children with relatively low IQs (around 80) from disadvantaged families (headed mostly by single, uneducated, and often unemployed mothers) at the age of 3–4 where randomly assigned to two groups: (i) a treatment group of about 60 children who were sent to a high-quality preschool and where their families received additional support from professionals at home and (ii) a control group (of about 60 children), where children and their families did not get any additional support. The individuals of both groups where then interviewed and tested several times over their life course (long-term RCT). The interesting finding was that the treatment group behaved differently from the control group even up to age 40. The members of the treatment group were more likely to be employed, get higher earnings, and were less dependent on social welfare. The Perry Preschool experiment therefore established a remarkable long-term causal effect for a very specific study population (children from extremely poor families in the USA). Of course, it is desirable to understand and generalize the results of such experimental studies as broadly as possible. However, it is doubtful whether the causal effect can be credibly generalized to other social groups and other countries. A recent cross-national comparative life course study (Blossfeld, Kulic, Skopek, & Triventi, 2017) reveals that early child care and education is very differently organized across Europe and the Anglophone societies. These differences engender a broad range of short- and long-term effect patterns for children over the life course. In general, more advantaged families sent their children more frequently and earlier to formal childcare (e.g., in Germany, Sweden, Finland, Russia, or Italy); and children from disadvantaged families could gain more than advantaged children from center-based care (e.g., in Germany, UK, Norway, Netherlands, and Ireland). However, the main finding of the Blossfeld *et al*. (2017) volume is that the relative gains of disadvantaged children were quite small compared to the huge cognitive achievement gaps among children from different socioeconomic origins. Thus, the conclusion of the cross-national comparative life course study is that all children profited from early childcare (elevator effect), but social inequalities in achievement

between children from social groups were only marginally reduced by early childcare. In other words, the causal effect of the Perry Preschool experiment cannot simply be generalized to other country settings and different social groups, at least not according to the evidence from the available cross-national comparative observational life course data. The Perry Preschool Program engendered too optimistic expectations.

Another interesting feature of the Perry Preschool example again is typical for experimental studies: It is quite unclear why there is a long-term effect of early child care at all. Some researchers speculated that there is a cognitive mechanism behind the established positive causal effect of the early intervention (Schweinhart *et al.*, 2005), while others have stressed that a noncognitive (or motivational) mechanism might have been at work (Heckman, Stixrud, & Urzua, 2006). Thus, the most interesting scientific question, which is of course also important for possible future interventions, was not answered by the Perry Preschool experiment. All what we know is that we have measured a strong causal effect, but based on this experiment, we do not know why this is the case and whether it has external validity. Establishing the effect of a cause in an experiment alone is therefore of limited value.

The Issue of Long-Term Effect Shapes of RCT

The Perry Preschool experiment is also instructive with regard to the question of how long a causal effect is effective. Figure 1 clearly shows that the treatment of the Perry Preschool Program worked and the IQ of the experimental group increased at the ages of 3 and 4. However, and that is not only surprising but also theoretically important, the causal effect on IQ afterward declined and eventually completely vanished at age 10. Thus, the causal effect on IQ dies gradually with increasing age. With regard to the age-related changing causal force of the treatment at least two interpretations are possible. First, the causal effect on IQ is only local and does not have any lasting impact over the life course. This would be a disappointing message for "no child left behind" intervention programs. The second interpretation is that the causal effect on IQ is still there, but that it is increasingly overwritten by even stronger causal forces connected with the school environment. Thus, if children of the treatment group are increasingly exposed to the everyday experiences of the standard school system in Ypislanti, the positive impact of the preschool intervention is increasingly lost. This hypothesis would suggest that the Perry Preschool Program needs to be continued into the primary school age or even beyond.

The central lesson to be learned from this time-varying causal effect is that the credibility of the results of the RCT depends on further causal factors that gain importance after the initial treatment but are not under control.

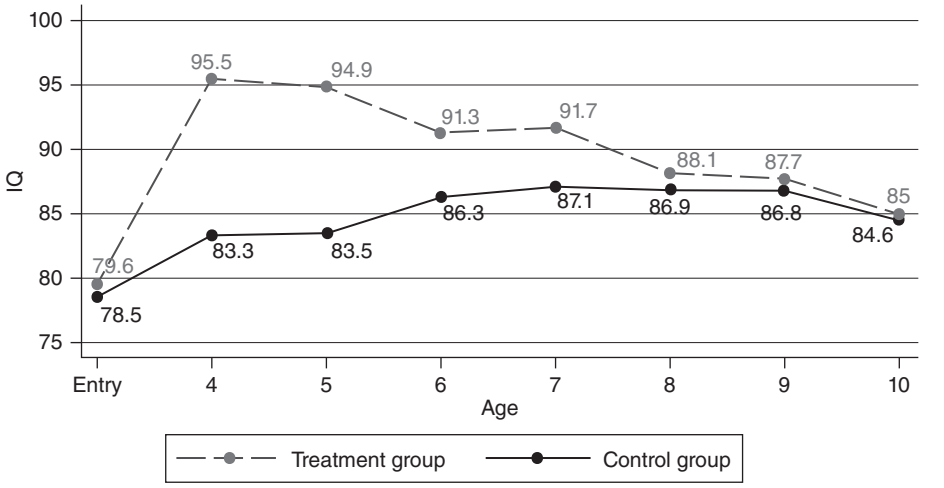

Figure 1 IQ development by age and treatment group in the Perry Preschool Program. (Test was administered at program entry and each of the ages indicated.) *Source*: Perry Preschool Program. IQ measured on the Stanford-Binet Intelligence Scale (Terman & Merrill, 1960).

The initial causal effect is undermined by causal influences of changing life course contexts and/or developmental processes. Of course, all these questions could in principle be studied with a series of complex experiments, but this would be a very expensive and time-consuming enterprise. In particular, this approach would need a theory when exactly causal effects should be measured and how they change in the life course. Life course studies based on (provisional) theory are more efficient in this respect, but of course, they have the problem that they only provide estimates with lower credibility.

AGENCY VERSUS RANDOMIZATION

In life course research, randomization is often practically or socially unacceptable. For example, it is morally and legally impossible to assign twins at birth randomly to different social origin families in order to measure the impact of different family environments on school success. In addition, strict experimental controls are often hard to apply in the life course. Thus, well-designed randomized controlled experiments are rarely applied by practitioners in life course research and most of this research is based on nonexperimental observations. Many of the life course processes are also only accessible with retrospective designs (e.g., see the German Life History Study, GLHS), precluding any experimental manipulations.

Indeed, randomization and manipulation might particularly conflict with one of the genuine research goals of a life course researcher. They typically

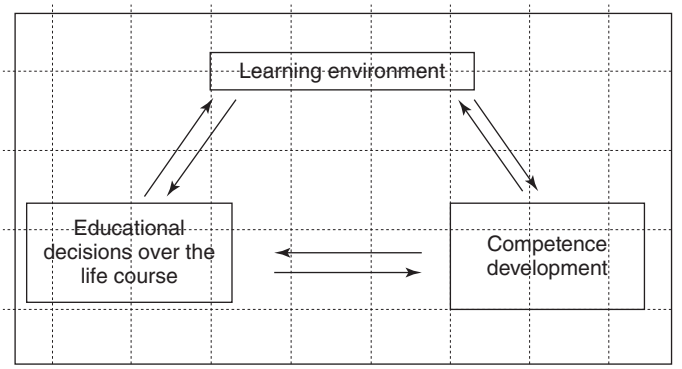

Figure 2 Educational decisions and the development of competencies throughout the life course.

want to understand and explain the selection processes over the life course. Subjects have agency, beliefs, and interests that influence their actions and choices. Thus, the life course engenders a dynamic interdependence (Figure 2) between (i) (self-) selection to a specific "treatment" (e.g., type of school), (ii) the exposure to the "treatment" itself (e.g., educational environment), (iii) the outcome from the "treatment" (e.g., academic ability), and (iv) the (self-) selection to the next "treatment" in the life course (e.g., entering university or not), which is of course again based on the previous outcome (the academic ability) and so on. In other words, this kind of life course logic creates reciprocal dependences among (self-) selection, treatment, and outcome (Figure 2). Blossfeld, Kilpi-Jakonen, Vono de Vilhena, and Buchholz (2014) demonstrate that the Matthew effect logic is particularly strong in the case of adult learning in many modern societies. In this case, randomization would simply destroy the cumulative life course logic. For example, in the German tracked school system, parents systematically select secondary schools, which influence their children's achievements differently, and these different outcomes lead to new educational choices in the next step (Blossfeld, Buchholz, Skopek, & Triventi, 2016). This means that children are rarely assigned by a lottery (as in the RCT) to a secondary school. However, note that this was the case in the Republic of Korea some years ago (Blossfeld, Blossfeld, & Blossfeld, 2017).

In the German tracking system, where children are basically (self-) selected to the academic or nonacademic track after age 10, randomization in an RCT would mean that an experimental control group would be created that would be composed of students who would have entered the academic track under normal life course conditions, but who were randomly denied access to this track in the RCT. Under the assumption that students, who would

enter into the academic and nonacademic tracks, do not differ in terms of learning gains (the conventional common effect model), randomization would be no problem. However, if the students differ in their academic abilities and motivations, randomization would alter the pool of students in the academic and the nonacademic tracks (as well as their behavior) and create what Heckman and Smith (1995, p. 92) have called a *randomization bias*. This means that the randomization would change the substantive process under study and therefore measure a peculiar ATE in the RCT.

Agency has also consequences for policy interventions that are based on RCTs. For example, when policy makers reform the educational system from a tracking system to a system of comprehensive schools in order to create more equality of educational opportunity, privileged parents might respond to this policy change. For example, they might create their own new private elite school, as it happened in Bremen (Germany) when a "Ökumenisches Gymnasium" was created by upper class parents in response to the citywide introduction of the comprehensive school. Thus, parents cannot be considered as passive human beings, but they have agency and behave strategically with regard to policy interventions. They want to do the best for their own children, and therefore, they try to escape the equalizing educational reform. Thus, a major issue of all educational reforms is the role of parent's decisions and strategic behavior (secondary effect of social origin). For example, it is well documented in educational research that if the influence of parents decision-making is reduced by teacher's compulsory school recommendations (which are more closely tied to academic achievement in school), equality of educational opportunity increases (Dollmann, 2016). In a new cross-national comparison of the effects of secondary school organizations, Blossfeld *et al*. (2016) also document analysis results from life course research which support this argument. They show that advantaged families strategically exploit different opportunities provided by different school organizations in order to achieve their goals. Of course, this agency mechanism limits the impact of educational reforms aiming to reduce inequalities of educational opportunity.

In sum, a serious issue for life course researchers arises from the insistence of the exponents of the "causation as consequential manipulation" approach that causes must be manipulable (by an experimenter—at least in principle) (Holland, 1986). The idea is that once the treatment or intervention is introduced, it will automatically lead to an outcome (stimulus–response model of behavior): $X_{tr} \rightarrow Y_{tr}$. The units of analysis in the social sciences, the individuals, are therefore assumed to be quite passive subjects whose behavior is explained only by causal factors and their " … 'objectives, knowledge, reasoning, and decisions' have no further relevance" (Goldthorpe, 2001, p. 8). This understanding of causation clearly excludes important process models

and reduces the testability of relevant theories and models for life course research.

QUASI-EXPERIMENTAL DESIGNS WITH OBSERVATIONAL DATA AND THEIR LIMITATIONS

Given the attractiveness of RCT for causal inference, social scientists have increasingly applied techniques that mimic RCT, so that they can more credibly assume that they have identified "causal effects" (Deaton & Cartwright, 2016; Sobel, 2005). Examples of those techniques are (i) panel studies where researchers use repeated individual measurements to adjust for time-invariant unobserved individual characteristics (fixed effects); (ii) models that include additional statistical terms to control for unobserved heterogeneity in hazard rate models; (iii) the inclusion of an instrumental variable (IV) that is correlated with the independent variable but not with the dependent variable, so that the impact of a confounding variable can be controlled for; (iv) utilizing observed characteristics and propensity scores to create matched samples, much as they would in a RCT setting; or (v) regression discontinuity models that compare individuals just above or below a crucial cutoff point where the individuals are likely to be quite similar with regard to a set of unobserved influences (Nicols, 2007; Schneider *et al*., 2007). The problem is that these quasi-experimental techniques for observational data often do not solve the selection and confounding problem of causal inference because they are often not applicable (e.g., there is no IV) or they make strong assumptions about the unobservables (Heckman, 2005), which are then untestable. Thus, how valuable these techniques might be, " … it is still difficult to avoid the conclusion that, in non-experimental social research, attempts to determine the effects of causes will lead not to results that 'never die' but only to ones that have differing degrees of plausibility. … (In other words/HPB), such results will have to be provisional in just the same way and for just the same reasons as those of attempts to determine the causes of effects via the 'partialling' approach." (Goldthorpe, 2001, p. 6) Thus, these quasi-experimental models offer the opportunity to estimate models under different statistical assumptions. As long as the estimation results of these models are consistent, this increases the credibility of the effects under different model conditions. However, if these models provide different or even contradicting estimations, the problem is which model is more credible. Since most models depend on untestable statistical assumptions, the choice between models is to some extend arbitrary. Thus, it seems that the benefits of the "causation as consequential manipulation" approach for life course research that works with observational data is quite limited.

CONCLUSIONS

Causal inference clearly should not be limited entirely to a matter of theory-free statistical predictability as in the causation as robust dependence approach. Well-designed controlled experiments are important study designs for causal inference, but since in life course research randomization is often not unacceptable or inappropriate, they are rarely applied by practitioners. Thus, most inferences in life course research have to be based on non-experimental observations of social processes. Under these conditions, both approaches, causation as consequential manipulation and causation as generative process, need to try to eliminate spurious (or non-causal) influences and will therefore never lead to results that "never die" but only to ones that have differing degrees of plausibility.

Furthermore, the approach of causation as consequential manipulation is too restrictive for modern life course research because the idea is that once the treatment or intervention is introduced, it will automatically lead to an outcome. The units of analysis in the social sciences, the individuals, are therefore assumed to be passive subjects whose behavior is explained only by causal factors. A necessary augmentation of the two understandings of causation is therefore the idea of causation as generative process, proposed by David Cox. According to this view, it is crucial to the claim of causation that there is an elaboration of an underlying (substantive) generative process existing in time and space. This argument will be developed further in the second part of the pair of essays (*see* Evidence of Causation—The Contribution of Life Course Research, Part II: Causation as Generative Process).

REFERENCES

Agresti, A., & Franklin, C. (2007). *Statistics. The art and science of learning from data*. Upper Saddle River, NJ: Person Prentice Hall.

Blalock, H. M. (Ed.) (1970). *Causal models in the social sciences*. Chicago, IL: Aldine.

Blau, P. M., Duncan, O. D., & Tyree, A. (1967). *The American occupational structure*. New York, NY: Free Press.

Blossfeld, P. N., Blossfeld, G. J., & Blossfeld, H.-P. (2017). The speed of educational expansion and changes in inequality of educational opportunity. In T. Eckert & B. Gniewosz (Eds.), *Bildungsgerechtigkeit* (pp. 77–92). Wiesbaden, Germany: Springer VS.

Blossfeld, H.-P., Buchholz, S., Skopek, J., & Triventi, M. (Eds.) (2016). *Models of secondary education and social inequality. An international comparison*. Cheltenham, England: Edward Elgar Publishing.

Blossfeld, H.-P., Golsch, K., & Rohwer, G. (2007). *Event history analysis with Stata*. Mahwah, NJ: Erlbaum.

Blossfeld, H.-P., Kilpi-Jakonen, E., Vono de Vilhena, D., & Buchholz, S. (Eds.) (2014). *Adult learning in modern societies: Patterns and consequences of participation from*

a life-course perspective. eduLIFE Lifelong Learning Series (Vol. 1). Cheltenham, England: Edward Elgar Publishing.

Blossfeld, H.-P., Kulic, N., Skopek, J., & Triventi, M. (Eds.) (2017). *Childcare, early education and social inequality – An international perspective*. eduLIFE Lifelong Learning Series (Vol. 4). Cheltenham, England: Edward Elgar Publishing.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Buchmann, M. (1989). *The script of life in modern society: Entry into adulthood in a changing world*. Chicago, IL: University of Chicago Press.

Cox, D. R. (1958). *Planning of experiments*. New York, NY: Wiley.

Deaton, A., & Cartwright, N. (2016). *Understanding and misunderstanding randomized controlled trails* (Working Paper 22595). Cambridge, MA: National Bureau of Economic Research.

Dollmann, J. (2016). Less choice, less inequality? A natural experiment on social and ethnic differences in educational decision-making. *European Sociological Review*, *32*, 203–215.

Duncan, O. D. (1966). Path analysis: Sociological examples. *American Journal of Sociology*, *72*, 1–16.

Elder, G. H., Jr., & Giele, J. Z. (2009). *The craft of life course research*. New York, NY: The Guilford Press.

Elder, G. H., Jr., Kirkpatrick Johnson, M., & Crosnoe, R. (2004). The emergence and development of life course theory. In J. T. Mortimer & M. J. Shanahan (Eds.), *Handbook of the life course* (pp. 3–19). New York, NY: Springer.

Elder, G. H., Jr., & Shanahan, M. J. (2006). The life course and human development. In R. E. Lerner & W. Damon (Eds.), *The handbook of child psychology, Vol. 1: Theoretical models of human development* (6th ed., pp. 665–715). New York, NY: Wiley.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, England: Oliver & Boyd.

Fox, J. (1992). Statistical models for nonexperimental data: A comment on Freedman. In J. Popper Shaffer (Ed.), *The role of models in nonexperimental social science: Two debates* (pp. 63–66). Washington, DC: American Educational Research Association and the American Statistical Association.

Freedman, D. A. (1992). As others see us: A case study in path analysis. In J. Popper Shaffer (Ed.), *The role of models in nonexperimental social science: Two debates* (pp. 3–30). Washington, DC: American Educational Research Association and the American Statistical Association.

Goldthorpe, J. H. (2001). Causation, statistics, and sociology. *European Sociological Review*, *17*, 1–20.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*, 424–438.

Heckman, J. J. (2005). Rejoinder: Response to Sobel. *Sociological Methodology*, *35*, 135–162.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives*, *9*, 85–110.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). *The effects of cognitive and noncognitive abilities on labor market outcomes and social behaviour* (Working Paper 12006). Cambridge, MA: National Bureau of economic research.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equation models. *Sociological Methodology*, *18*, 449–484.

Jackson, M., & Cox, D. R. (2013). The principles of experimental design and their application in sociology. *Annual Review of Sociology*, *39*, 27–49.

Johnston, J. (1972). *Econometric methods* (2nd ed.). New York, NY: McGraw-Hill.

Lieberson, S. (1985). *Making it count. The improvement of social research and theory*. Berkeley: University of California Press.

Marini, M. M., & Singer, B. (1988). Causality in the social sciences. In C. C. Clogg (Ed.), *Sociological methodology* (pp. 347–409). Washington, DC: American Sociological Association.

Mayer, K. U. (1997). Notes on a comparative political economy of life courses. *Comparative Social Research*, *16*, 203–226.

Mayer, K. U. (2009). New dimensions in life course research. *Annual Review of Sociology*, *35*, 413–433.

Mayer, K. U. (2015a). An observatory for life courses: Populations, countries, institutions, and history. *Research in Human Development*, *12*, 196–201.

Mayer, K. U. (2015b). The German Life History Study – An introduction. *European Sociological Review*, *31*, 137–143.

Nicols, A. (2007). Causal inference with observational data. *The Stata Journal*, *7*, 507–541.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*, 237–257.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, *6*, 34–58.

Rubin, D. B. (1980). Discussion of 'randomization analysis of experimental data in the Fisher randomization test' by Basu. *Journal of the American Statistical Association*, *81*, 961–962.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects. Using experimental and observational design*. Washington, DC: American Educational Research Association.

Schweinhart, L., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). *Lifetime effects: The High/Scope Perry Preschool Study through age 40* (pp. 194–215). Ypsilanti, MI: High/Scope Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shanahan, M. J., Mortimer, J. T., & Kirkpatrick Johnson, M. (2016). *Handbook of the life course* (Vol. *II*). Cham, Switzerland: Springer.

Sobel, M. E. (1995). Causal inference in the social and behavioral sciences. In G. Arminger, C. C. Clogg & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 1–38). New York, NY: Plenum Press.

Sobel, M. E. (2005). Discussion 'The scientific model of causality'. *Sociological Methodology*, *35*, 99–133.

Sørensen, A. B. (2009). Statistical models and mechanisms of social change. In P. Hedström & B. Wittrock (Eds.), *Frontiers of sociology* (pp. 369–400). Leiden, The Netherlands: Brill.

Terman, L. M., & Merrill, M. A. (1960). *Stanford–Binet intelligence scale: Manual for the third revision form L–M with revised IQ tables by Samuel R. Pinneau*. Boston, MA: Houghton Mifflin.

HANS-PETER BLOSSFELD SHORT BIOGRAPHY

Hans-Peter Blossfeld is on leave from his Chair in Sociology at the Faculty for Social Sciences, Economics, and Business Administration at the University of Bamberg, Germany, until August 2017. Since 2012 he has a Chair in Sociology at the European University Institute in Florence, Italy. He has published 45 books and over 300 articles on life course research, social inequality, youth, family, educational sociology, labor market research, demography, social stratification and mobility, the modern methods of quantitative social research, and statistical methods for longitudinal data.

RELATED ESSAYS

Learning Across the Life Course *(Sociology)*, Jutta Allmendinger and Marcel Helbig

Repeated Cross-Sections in Survey Data *(Methods)*, Henry E. Brady and Richard Johnston

Genetics and the Life Course *(Sociology)*, Evan Charney

Patterns of Attachments across the Lifespan *(Psychology)*, Robyn Fivush and Theodore E. A. Waters

Lifecourse and Aging *(Anthropology)*, Haim Hazan

Social Inequality across the Life Course: Societal Unfolding and Individual Agency *(Psychology)*, Jutta Heckhausen

Meta-Analysis *(Methods)*, Larry V. Hedges and Martyna Citkowicz

Maternal and Paternal Employment across the Life Course *(Sociology)*, Michaela Kreyenfeld

From Individual Rationality to Socially Embedded Self-Regulation *(Sociology)*, Siegwart Lindenberg

Longitudinal Data Analysis *(Methods)*, Todd D. Little *et al*.

Social Change and Entry to Adulthood *(Sociology)*, Jeylan T. Mortimer

Schooling, Learning, and the Life Course *(Education)*, Aaron M. Pallas

Modeling Life Course Structure: The Triple Helix *(Sociology)*, Tom Schuller

Gender and the Transition to Adulthood: A Diverse Pathways View

(Sociology), Ingrid Schoon
Becoming Adult: Meanings of Markers to Adulthood *(Sociology)*, Richard A. Settersten, Jr. *et al.*
Causation, Theory, and Policy in the Social Sciences *(Sociology)*, Mark C. Stafford and Daniel P. Mears
Crime and the Life Course *(Sociology)*, Mark Warr and Carmen Gutierrez
The Gendered Transition to Parenthood: Lasting Inequalities in the Home and in the Labor Market *(Sociology)*, Marie Evertsson and Katarina Boye