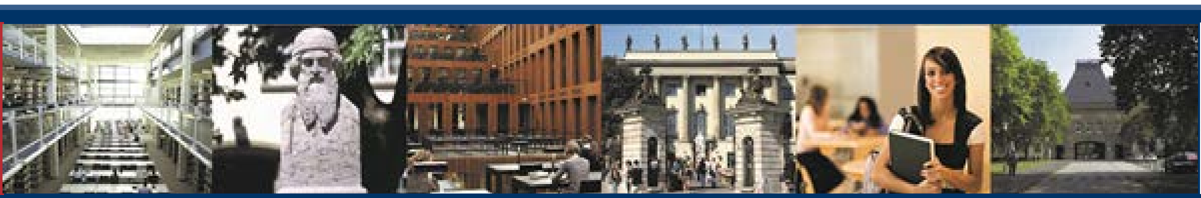

Modeling and Measuring
Competencies in Higher Education
Http://www.kompetenzen-im-hochschulsektor.de/index_ENG.php

KoKoHs Working Papers

No. 4

Stefanie Berger, Svenja Hammer, Stefan Hartmann,
Cora Joachim & Thomas Lösch

Causal Inference in Educational Research

Approaches, Assumptions and Limitations

SPONSORED BY THE

Federal Ministry
of Education
and Research

Johannes Gutenberg University Mainz

Humboldt University of Berlin

KoKoHs Working Papers from the BMBF-funded research initiative „Modeling and Measuring Competencies in Higher Education“

The *KoKoHs Working Papers* series publishes articles from the funding initiative „Modeling and Measuring Competencies in Higher Education (KoKoHs)“. These may be conceptual papers or preliminary results intended for rapid dissemination or public discussion. Publication in *KoKoHs Working Papers* does not preclude publication of the texts elsewhere. The responsibility for the content lies with the authors. The content does not necessarily reflect the views of the publishers of *KoKoHs Working Papers*.

Publishers:

Prof. Dr. Sigrid Blömeke
Humboldt University of Berlin
Faculty of Arts and Humanities IV
Department of Education Studies
Chair of Instructional Research
Unter den Linden 6
D-10099 Berlin

Prof. Dr. Olga Zlatkin-Troitschanskaia
Johannes Gutenberg University Mainz
Department 03: Law, Management and Economics
Chair of Business Education I
Jakob Welder-Weg 9
D-55099 Mainz

Contact:

christiane.kuhn@uni-mainz.de
corinna.lautenbach@hu-berlin.de

The research initiative is funded by the German Federal Ministry of Education and Research under grant no. 01PK11100A and 01PK11100B.

The *KoKoHs Working Papers* are also available for download:

http://www.kompetenzen-im-hochschulsektor.de/index_ENG.php

Causal Inference in Educational Research

Approaches, Assumptions and Limitations

Stefanie Berger
Svenja Hammer
Stefan Hartmann
Cora Joachim
Thomas Lösch

Contact:

stefanie.berger@bwl.uni-mannheim.de
svenja.hammer@leuphana.de
stefan.hartmann@hu-berlin.de
cora.joachim@biologie.uni-goettingen.de
thomas.loesch@uni-tuebingen.de

Bibliographical references:

Berger, S., Hammer, S., Hartmann, S., Joachim, C., & Lösch, T. (2013). Causal Inference in Educational Research. Approaches, Assumptions and Limitations. (KoKoHs Working Papers, 4). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.

Causal Inference in Educational Research – Approaches, Assumptions and Limitations

Abstract:
The Working Paper gives an overview about the topic of causal inference, covered in the Institute on Statistical Analysis for Education Policy organized by the American Educational Research Association in spring 2013. Because randomized experiments are often difficult to implement into large-scale studies in educational research, inference on the causality of different treatments (e.g. different teaching approaches) is limited. This paper discusses the possibilities to draw causal inferences in non-randomized experiments, and provides an introduction to different analytical approaches, namely propensity score analysis, sensitivity analysis and the regression discontinuity approach. The main idea behind each procedure is explained, advantages and limitations are discussed.

Keywords:
Causal inference, propensity score, sensitivity analysis, regression discontinuity

Contents

Introduction

In educational studies, researchers are often interested in drawing causal inferences, such as: "Does program A increase the student's achievement in Mathematics more than program B?" However, the apparent success of a program might be caused by other factors than the elements of the program itself and it is simply not possible to account for all causes of observed effects and to identify the relation between these causes (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). Thus, it is impossible to definitely determine the effect that the programs had on participating students. Randomized experiments provide one possibility to overcome this problem. However, randomization is often not possible in educational settings. Instead, researchers often apply analytic approaches, e.g. propensity scores, regression discontinuity designs or sensitivity analyses to estimate causal effects.

As causal inference is a central issue in educational research, the American Educational Research Association (AERA) organized a statistics institute about this topic in spring 2013. This paper gives an overview about the issues covered in the institute. First, the causal models of Campbell and Rubin will be introduced. We then give an overview of different methods for estimating causal effects. For deeper insights, we provide recommendations for further reading.

1. Models of Causality

Donald T. Campbell and Donald Rubin proposed different models of causal relationships and their magnitude. Both authors studied effects of known causes instead of the causes of known effects (Shadish, 2010). The models will be explained in the following chapters.

Campbell's causal model

Campbell's causal model starts with the identification of two categories of inferences that scientists make:

1) Whether the stimulus caused an effect (internal validity) and
2) to what extent the cause-effect relation can be generalized (external validity) (Shadish, 2010).

Campbell later differentiated between statistical conclusion validity, internal validity, construct validity and external validity (Shadish, 2010). In order to enable correct inference from experiments, it is crucial that validity is asserted. Threats to validity, also called alternative explanations, have to be avoided, since they question drawn inference. For instance, it should be assured that an effect is caused by the treatment and not by confounding variables (Shadish, 2010). Table 1 provides exemplary threats to validity.

exemplary threats to validity	
internal validity	*external validity*
participants mature during the period of the study	differences between treatment and control group
subjects drop out during the study	pre-test influences the participants' performance in the test

Tab. 1: Exemplary threats to validity.

The design of any study should be analyzed with regards to validity and threats to validity, aiming to minimize threats to validity. Internal validity is most important since a) the effect of the treatment is crucial for the study and b) it is the basis for generalizations (Shadish, 2010).

Rubin's causal model

The approach of Rubin is to look for effects of given causes. For instance, a researcher can study the effects different programs have on children's achievement (Schneider et al., 2007). If the effects of the programs differ (e.g. one group reaches higher achievements) and the two groups of children are comparable in all respects, then the researcher can conclude that one program is more suitable to increase the achievement (Schneider et al., 2007).

A first approach from Rubin and his colleagues to causal inference was to assume the ideal but impossible case: An individual receives a treatment A and the effects of the treatment are measured. Then time is turned back and the same individual receives treatment B (the control group treatment). The causal effect is the difference in the effects of the two treatments (Schneider et al., 2007).

In Rubin's model, causal inference is explained as follows: A unit (u), for instance a person, gets one of two different treatments (t, c). Treatment variables in experiments are manipulated by the researcher; in that sense, given variables like race or gender cannot be treatment variables. Y_t and Y_c are the corresponding responses to the treatments. The causal effect is Y_t-Y_c. The problem of causal inference is that only Y_t or Y_c can be observed, since each unit can get one treatment only (Holland, 1986; Shadish, 2010). As a consequence, it is important that the units are assigned randomly to the treatment.

The "randomized controlled experiment is the most powerful design for detecting treatment effects" (Schneider at al., 2007). However, it is not always possible to conduct a randomized controlled experiment due to logistical, financial or ethical reasons. For instance, children cannot be assigned to schools randomly. A potential alternative solution is the creation of large-scale datasets that approximate a randomized experiment (Schneider et al., 2007).

2. Methods for estimating causal effects

When working with observational data, the following methods can be used to approximate randomized controlled experiments (Schneider et al. 2007, p. 42):

- Propensity Score Modeling
- Regression Discontinuity
- Sensitivity Analysis
- Fixed Effects Models
- Instrumental Variables.

In the following sections, we describe the first three methods in detail. This article, however, will not go into detail regarding fixed effects models and instrumental variables.

2.1 Propensity Score Modeling

Estimating the causal effect of a treatment is essentially a problem of missing data. As every subject in the study either does (treatment group) or does not receive the treatment (control group), but never both, we can only compare the outcome of the treatment group with the outcome of the control group. However, the effect a treatment would have had on a subject in the control group remains unknown. The results of the two groups may be compared directly when randomized experiments are used, as the subjects in the two groups are then expected to be similar. In non-randomized experiments, however, there is a high risk that the subjects in the treatment group are systematically different from the subjects in the control group and a direct comparison may yield misleading results. One way to overcome this problem and to make comparisons more meaningful is to group treated and non-treated subjects by using propensity scores (Rosenbaum & Rubin, 1983).

Propensity scores quantify the probability that an individual is assigned to the treatment group when assignment is not randomized. The propensity score can be formally defined as

$$e(x) = \Pr(t= 1|x),$$

where $\Pr(t= 1|x)$ is the conditional probability of receiving a treatment t, given covariates x (Rosenbaum & Rubin, 1983, p. 42). In contrast to randomized experiments, in which only one specification of the propensity score function e(x) exists, e(x) is unknown in nonrandomized trials but can be estimated from observed data, e.g. by using a logit model (Rosenbaum & Rubin, 1983). Thereby, the aim is to control for all variables that likely predict group membership and that are expected to be related to the outcome (Shadish, Cook, & Campbell, 2002). The necessary calculations can be carried out using statistical software, such as SPSS, SAS, or R.

Approaches using propensity scores

Propensity scores can be applied in different ways before the treatment effect is estimated. The most popular approaches are (Frank et al., 2008; Rosenbaum & Rubin, 1983):

(1) Propensity score matching,
(2) Propensity score stratification, and
(3) Inverse-probability-of-treatment weighting.

Propensity score matching aims to “produce a control group of modest size in which the distribution of covariates is similar to the distribution in the treated group” (Rosenbaum & Rubin, 1983, p. 48). In essence, one looks for *twins* in the data set, where one twin is in the control group and the other is in the treatment group. By calculating propensity scores, each subject’s set of covariates is reduced to a single score. Thus, when applying propensity score matching, one simultaneously accounts for multiple variables (Shadish, Cook, & Campbell, 2002). In a best-case scenario, subjects are exactly matched with regard to all covariates x. Then, an identical distribution of x would be obtained for both the treatment and control group. However, in reality we frequently have to use approximations as identical matches are rarely possible (Rosenbaum & Rubin, 1983). Since the distribution of x in the two groups is more similar in matched samples than in random samples, the variance of the estimated average effect of the treatment is lower when matching is applied than when the sample is randomized (Rosenbaum & Rubin, 1983). An example for this method is given in the next section.

In the second standard approach, subjects are divided into strata or subclasses with similar characteristics. The matching process is based on an algorithm that minimizes “aggregate sample differences between treatment and control conditions on the propensity score” (Shadish, Cook, & Campbell, 2002, p. 162). Typically, the strata are formed according to quintiles; a total number of five subclasses is fairly common (Shadish, Cook, & Campbell, 2002). As each subject in a subclass has an equal probability of receiving the treatment, this approach approximates random sampling (Schneider et al., 2007). Therefore, stratification allows for direct comparison between the two groups within each stratum (Rosenbaum & Rubin, 1983).

In the third major approach, each subject is weighted by his or her propensity of being in the treatment group. Thus, different propensities of group assignments are adjusted for by using corresponding weights. We exemplify this method in the following section (Frank et al., 2008).

Example

Frank et al. (2008) use sociometric data to analyze whether the American teacher certification process by the National Board of Professional Teacher Standards (NBPTS) has an effect on teachers' readiness to help their colleagues. As board certification happens on a voluntary basis, it is not possible to assign teachers randomly to the treatment and control groups. Thus, the authors use propensity scores to compare certified and non-certified teachers. Two different ways to use propensity scores are discussed.

First, one could apply propensity score matching (Method 1) and assign each NBPTS certified teacher to another teacher that is not board certified but has a similar propensity of becoming certified. Then analyses regarding the differences in outcomes between the matched pairs could be made. The problem with this method is that a considerable amount of data is lost because it is only possible to analyze those teachers that received the treatment and their matches. All teachers that are not board certified and not matched would be removed from the data set.

A second way to apply propensity scores in this example is to regress the outcome of interest on treatment, whereby teachers are weighted by their propensity scores. Teachers who receive the treatment are weighted by $\frac{1}{\mathrm{e(x)}}$, whereas teachers in the control group are weighted by $\frac{1}{1-\mathrm{e(x)}}$.

Thus, the statistical weight of certified teachers increases when the propensity of being certified is decreasing. Conversely, the statistical weight of non-certified teachers increases with an increasing propensity of being certified. Therefore, the analysis focuses on the comparison between those teachers who are certified but had a low propensity and those teachers who are not certified yet had a high propensity. Heckman (2005) refers to this estimate as "the effect of the treatment for people at the margin of indifference (EOTM)" (as cited in Frank et al., 2008, p. 20).

In studies focusing on treatments that are relevant for policy decisions, the interpretation of this calculation is especially relevant. If policy makers are interested in changes in incentives for those who receive a treatment (e. g. to become board certified), then estimates of effects should concentrate on those subjects who are most likely to respond to policy changes. This corresponds to those who opted into the treatment group but have low propensity for doing so. Thus, they might not have received the treatment if there were fewer incentives (Frank et al., 2008, p. 19). One would also be interested in those subjects who are in the control group but have high propensity for being in the treatment group and therefore might respond to increases in incentives.

Advantages and limitations

Propensity scores are a useful tool for estimating the effect of a treatment when randomized experiments are not possible, or when subjects "self-selected themselves into treatment and control groups" (Schneider et al., 2007, p. 50). By applying propensity scores, one can correct for biases from observed characteristics. Essentially, propensity score approaches are "a version of regression or matching that allows researchers to focus on the observed covariates that "matter most'" (Schneider et al., 2007, p. 49). Rather than carrying out a regression analysis for the whole sample in non-randomized studies, propensity scores are used to evaluate whether the covariates are balanced between the treatment and the control group. This balance, in turn, supports causal inference, because it allows for adjustment of potential bias attributed to the observed variables (Frank et al., 2008).

Despite the advantages of propensity score methods, a number of limitations should be considered. First of all, many non-randomized experiments have small sample sizes, but propensity scores work best with large samples. Additionally, as the example above demonstrates, the sample sizes can be limited by matching or stratifying when the overlap between the groups is low (Shadish, Cook, & Campbell, 2002). Furthermore, propensity scores only control for biases from *observed covariates*, while other methods (e. g. instrumental variables, fixed effect models) can also correct for omitted variable bias. Hence, the method assumes that no other confounding variables exist that are related to the propensity of group selection and to the outcome. However, the cost of measuring some variables might be prohibitively high, and even in cases in which the propensity score is calculated very carefully, it is impossible to account for all conceivably related variables. Thus, a certain amount of hidden bias still remains (Schneider et al., 2007; Shadish, Cook, & Campbell, 2002). To evaluate if the effect would be robust to hidden biases, propensity score matching can be supplemented by sensitivity analyses (Shadish, Cook, & Campbell, 2002).

Practical application

Propensity score analyses can be conducted using recent statistical software. Stuart (2012) gives a helpful overview concerning available packages and documentations for R, Stata, SAS, and SPSS. In the following, we will describe propensity score matching using the R package *MatchIt* (Ho, Imai, King, & Stuart, 2011). Only a few commands are necessary to utilize several matching methods. Stratification and weighting are also possible, but not intended by the developers and therefore require more complex programming. However, for users that are experienced with logistic regression these approaches readily applicable.

We assume a data frame that contains the variable columns y (criterion), x (treatment), as well as a, b, and c (covariates). We want to match the subjects in the treatment group (x = 1) to the subjects in the control group (x = 0) regarding their covariates. Finally, we want to test if the matched groups differ in y.

First, we match both groups using the “nearest neighbor” method: each subject in the treatment group is matched to the subject in the control group that has the most similar propensity score.

```
match <- matchit(x ~ a + b + c, data=dat, method="nearest")
```

The object `match` (that is no data frame yet) can descriptively be checked for balance of covariates in the matched groups either using one or both of

```
summary(match)
plot(match)
```

To produce a new data frame consisting only of matched subjects and including the propensity score (in the column `distance`), we use the following command:

```
dat.matched <- match.data(match)
```

Now, we can estimate the treatment effect using standard methods, e. g. linear regression:

```
lm(y ~ x, data = dat.matched)
```

The regression weight can be interpreted as the causal treatment effect.

Recommendations for further reading
For more information concerning other matching procedures or estimations of the causal effect, Ho et al. (2011) give a detailed overview. Examples for the application of propensity scores can be found in Becker, Lüdtke, Trautwein, Köller, and Baumert (2012); Frank et al. (2008) and Jackson, Thoemmes, Jonkmann, Lüdtke, and Trautwein (2012).

2.2 Regression Discontinuity

Regression discontinuity design (abbreviated RDD) is a method to estimate a treatment effect in a non-randomized study. The usual approach can be administered if there is

- *a single outcome or criterion*,
- *two groups* differing concerning their treatment and
- a *single predictor determining the group* individuals are selected to.

There has to be a predefined cut-off in the criterion determining the group membership. An example situation would be to test if vocabulary training improves low achieving students' performance. The outcome is the performance in a vocabulary test at the end of a semester. The two groups are a treatment group receiving the training, and a control group (for example receiving mathematics training). The predictor for selecting the individuals is a vocabulary pretest. The cut-off could be a score of grade F in the pre-test.

RDD is different from a randomized experiment in such as there is a predefined selection criterion that determines if an individual receives a treatment. The main idea of this approach is that individuals close to the cut-off are almost identical concerning the predictor. Therefore, the only difference between the two groups close to the cut-off is that one group receives the respective treatment. In RDD, the difference in the outcome between these two groups close to the cut-off is used as causal effect of the treatment. Accordingly, there is another difference to randomized experiments: While in randomized experiments an average treatment effect across the *whole sample* is estimated (by using t-tests or ANOVAs), in RDD this is not necessarily true. Without further assumptions, the average treatment effect *at the cut-off* is estimated and used as causal effect (Bloom, 2009).

The general procedure is to estimate a regression that differs with regard to the treatment, and then to calculate the difference in the predicted score for the criterion at the cut-off. Therefore, the procedure is called regression *dis*continuity. This general procedure is illustrated in Figure 1.

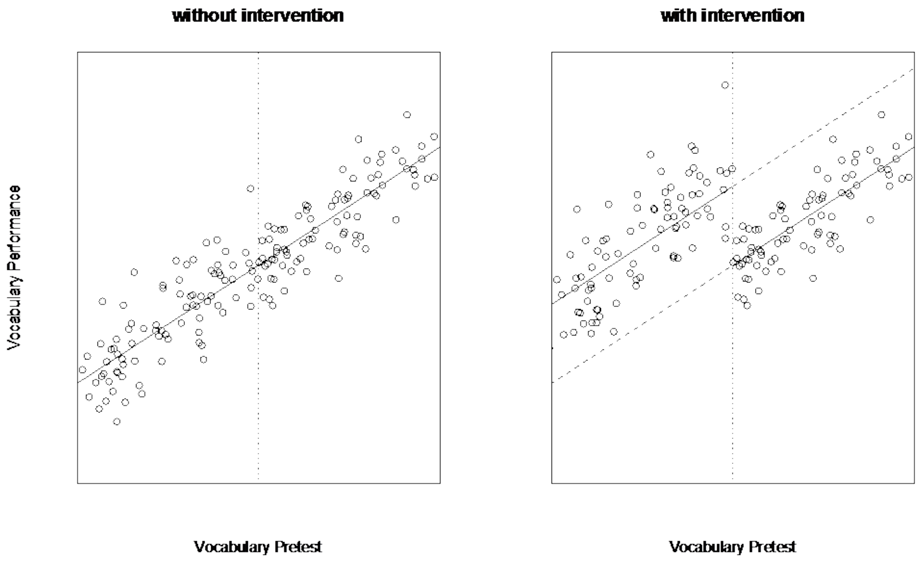

Figure 1: Hypothetical example of a regression discontinuity: estimated regression with and without treatment. Solid lines are “real” regressions; dashed lines are the counterfactual regressions.

On the left side, the relation between the vocabulary pretest and the following performance in the vocabulary test is shown. This would be the relation if no intervention would take place. The dashed line in the middle marks the cut-off, the solid line is the estimated linear regression. On the right the relation between the vocabulary pretest and the vocabulary test at the end of the semester is shown. The difference is that the individuals below the cut-off received the vocabulary training. Apparently, the vocabulary training improved the students’ performance. The dashed lines represent the hypothetical (or, as it is usually called, *counterfactual*) performance of the students if they had been in the respective other group.

Different Regression methods for RDD

To ensure that a RDD yields a precise causal effect, it is necessary to correctly specify the relationship between the predictor and the criterion. If this estimation is specified correctly, a RDD can produce a good estimate of a causal effect. There are several approaches to specify this relationship:

- Parametric regression (with linear regression as shown in Figure 1 as a special case)
- Non-parametric regression
- Semi-parametric regression

Parametric regressions are the most straightforward procedures, as they are relatively well known. They comprise linear, quadratic, or cubic relations between predictor and criterion.

Non-parametric regressions are mainly driven by the data. There are several approaches to estimate a smooth function based on the observations, which is not restricted by a parametric form. One approach is kernel regression. In a kernel regression, the predicted values are based on a linear combination of nearest-neighbor data points. Two decisions have to be made: how many neighbors are included (i. e. the *bandwidth*) and how the neighbors are weighted (i. e. the *kernel function*).

Semi-parametric models combine both approaches. A semi-parametric function contains parametric (for example: linear) and non-parametric elements.

In contrast to a standard regression analysis, the elements of the regression function are not of interest. It is important to specify the relation between the predictor and criterion correctly to obtain a valid estimate of the treatment's causal effect.

Evaluation of the method

The major advantage of RDD is that causal inference can be drawn in a non-experimental setting. To test if a RDD and a randomized experiment would be comparable concerning their results, Shadish, Galindo, Wong, Steiner, and Cook (2011) compared the two selection methods in a randomized experiment. In general, randomized experiments and RDD led to similar results concerning the direction of the effect. Nevertheless, the effect sizes were significantly higher in the RDD design than in the randomized experiment. However, the interpretation of this finding is not straightforward as both methods estimate different causal effects: a randomized experiment measures the average causal effect *in the whole sample* and the RDD measures the average causal effect *at the cut-off*.

Shadish (2011) proposed the following guidelines on how to apply an RDD to obtain unbiased estimates: (1) reduce self-selection of participants as much as possible, (2) obtain as much potential confounding variables as possible to control for their influence, and (3) use large samples.

Practical Application

In general, RDDs can be performed using diverse statistical software including *Stata* or *R*. We describe the basic procedure to perform an RDD in R. Understanding of the commands and the output requires a basic knowledge of R, the familiarity with the term *bandwidth*, and the knowledge of different types of *kernel functions* available.

An RDD can be performed using the package *rdd* (Dimmery, 2013). Under the assumption that the data frame `dat` has the column `x` containing the predictor variable and the column `y` containing the criterion, the function

```
rd <- RDestimate(y ~ x ,data=dat)
```

estimates a RDD and saves it in the object `rd`. The output appears as follows:

```
Call:
RDestimate(formula = y ~ x, data = dat)

Coefficients:
     LATE     Half-BW   Double-BW
-0.071228  -0.001703  -0.046338
```

The function `RDestimate()` fits a kernel regression model with a *triangular* kernel function and an automatically estimated bandwidth. Based on this model estimation, a causal effect is computed. The important part is the LATE, which is the estimation of the *local average treatment effect* (i. e. the causal effect). With `summary(rd)` a test for significance can be obtained. The user can specify the kernel function and the bandwidth using the arguments `bw = bandwidth` and `kernel = "kernel function"`, whereas `bandwidth` is a number and `kernel function` is a string specifying the kernel function. Finally, a RDD can be visualized using the `plot(rd)` function.

Note that using the rdd package, only kernel regressions can be fitted. Fitting of other parametric, semi-parametric or non-parametric models is not supported. A complete example (using historical data) can be found in Dimmery (2012).

Recommendations for further reading

The basic procedure on how to estimate a RDD is described in Bloom (2009). Teknomo (2007) describes kernel regression in a hands-on tutorial using an excel sheet. In Fahrmeier, Kneib, and Lang (2007), regression models are described, including non- and semi-parametric regression. This approach for kernel regression can be found under the (German) headline "Lokale Glättungsverfahren" (p. 333). Shadish (2011) gives general helpful advice for estimating causal effects under the condition of non-randomization.

2.3 Sensitivity Analysis

A different approach of dealing with causal inference in studies, in which random design is impractical or impossible, is sensitivity analysis (Frank, 2000). Not to be confused with sensitivity *power* analysis (Faul, Erdfelder, Lang, & Buchner, 2004), sensitivity analysis is a method to answer the question how robust a certain causal inference is to (unknown) confounding effects. In other words, sensitivity analysis assists a researcher in estimating how large a confounding effect would have to be to falsify causal inference from an analysis.

Confounding effects

Confounding variables are variables that correlate with both the predictor (x) and the outcome (y) of an analysis. This correlation can affect the results of the analysis and therefore bias the resulting (causal) interpretation. For example, a strong relation between migration background and test performance can be found in many educational studies. Students with a migration background score lower than students whose parents are natives (Stanat & Christensen, 2006). If we assume that the "student's test score is an accurate reflection of his or her mastery of a particular content area" (Martiniello, 2009, p. 161), we could draw a causal inference between descent (x) and competency in the tested content area (y) from this result, and conclude that migrants are less competent than natives. But of course, there are other variables to explain this phenomenon: For example, in any text-based assessments, reading abilities are needed to understand the test questions. This simple fact can constitute a language barrier for migrant students (National Research Council, 2000) and therefore bias test results (Hartmann, 2013).

As long as we know how much a certain confounding variable (like reading ability in our example) is statistically related to the predictor of interest (migration background) and to the outcome variable (test score), we can easily control for the confounding effect by including the confounding variable(s) as covariate(s) in the quantitative model. However, in scientific practice we often deal with "circumstances under which it is not possible to measure or control for the potentially confounding variable" (Frank, 2000, p. 147). Possible reasons are that the confounding variables are not accessible to the researcher, or that it is impossible or unethical to assign people randomly to these variables. However, it is possible to assess the robustness of a causal inference to the impact of any unknown or unmeasured confounding variable(s). Knowing about the robustness of the inference helps to defend the conclusion: The more robust the inference is, the more reasonable it is to interpret the predictor (x) as a true indicative of the causal effect.

Calculation

Assume a standard linear model for a dependent variable (y), a predictor variable (x), and an error (e):

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad \text{(equation 1)}$$

If the estimate of β_1 is statistically significant in equation 1 (i.e., the t ratio is larger than the critical t value for a given level of significance and degrees of freedom), we interpret it as indicative of an effect of x on y.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 cv_i + e_i \qquad \text{(equation 2)}$$

Equation 2 takes a confounding variable (cv) into consideration, which is correlated to both x and y. To assess the impact of cv on the inference of β_1, we ask "how large must the correlations be between cv and y and between cv and x such that β_1 is not statistically significant" (Frank, 2000, p. 152). To answer that question, Frank provides a statistical framework that boils down to three simple steps:

The first step is to calculate the correlation ($r_{x,y}$) between the predictor of interest (x) and the outcome (y) by partialling out all *known* covariates. We can perform this calculation with any statistical software, like SPSS, SAS, or R.

In the second step, a threshold value ($r^{\#}$) for r is defined such that r is statistically significant:

$$r^{\#} = \frac{t_{critical}}{\sqrt{(n-q-1)+t^2_{critical}}} \qquad \text{(equation 3)}$$

n = sample size
q = number of parameters estimated

In the third step, the impact (k) of any possible confounding variable is defined as the product of correlations $r_{x,cv}$ and $r_{y,cv}$ (assuming that $r_{x,cv} = r_{y,cv}$ to maximize the impact). Equation 4 illustrates the impact of k on the initial correlation.

$$r_{x,y|cv} = \frac{r_{x,y}-k}{1-k} \qquad \text{(equation 4)}$$

Finally, $r_{x,y|cv}$ is set to $r^{\#}$ and the equation is solved for k. We obtain the threshold of the impact of a confounding variable (ITCV) that would invalidate the inference:

$$ITCV = \frac{r_{x,y}-r^{\#}}{1-|r^{\#}|} \qquad \text{(equation 5)}$$

Even if we have no knowledge of the identity of the confounding variable, knowing the threshold enables discussion of the robustness of the initial effect of the known predictor (x) on the outcome (y). We can do so by simply putting the value for ITCV into relation to the strongest impact of all *known* covariates. In discussions, we can ask if theory provides any suggestions for variables that may have an impact equal or larger than the value for ITCV.

Recommendations for further reading
On his website, professor Kenneth Frank provides a Microsoft Excel spreadsheet which automatically calculates ITCV for any given t_{crit}, n, and observed t (Frank, 2013).

3. Summary

The methods presented in this working paper can be used to draw causal inferences in non-randomized studies. Table 2 shows an overview of the methods presented above.

	Propensity Score Matching	Propensity score Stratification	Inverse-probability-of-treatment Weighting	Regression Discontinuity	Sensitivity Analysis
Basic problem	Non-randomized samples				
When to use	- to achieve comparability between the treatment and the control group by accounting for each person's probability to be in one of those groups			- a single outcome or criterion - two groups differing concerning their treatment - a single predictor determining the group individuals are selected to	- to evaluate how robust an effect is against hidden biases
Advantages/ Disadvantages	- loss of data because non-matched persons are no longer viewed - propensity scores only control for bias from *observed covariates* - certain amount of hidden bias still remains - large samples required			- RDD is considered to have the highest internal validity (the ability to identify causal relationships in this research setting) - external validity may be lower, as the estimated treatment effect is local to the discontinuity - large samples required	
Procedure	- producing a control group of modest size that has (nearly) the same distribution of characteristics/ covariates than the treated group	- dividing persons with similar characteristics into strata or subclasses (typically into quintiles) - each subject in one subclass has an equal probability of receiving the treatment - comparison between the two groups within each stratum is possible	- each person is weighted by his or her propensity of being in the treatment group	- pretest-posttest two group design - pretest is used to assign persons to the treatment group (The main idea of this approach is, that individuals close to the cut-off are almost identical concerning the predictor) -the choice of cut-off value is usually based on one of two factors	- calculating the correlation between the predictor of interest and the outcome, by partialling out all *known* covariates - defining a threshold value - defining the impact of any possible confounding variable as a product of the correlations - helps to discuss the robustness of the initial effect of the known predictor on the outcome

Tab. 2: Methods for estimating causal effects.

References

Becker, M., Lüdtke, O., Trautwein, U., Köller, O., & Baumert, J. (2012). The differential effects of school tracking on psychometric intelligence: Do academic-track schools make students smarter? *Journal of Educational Psychology, 104*(3), 682-699.

Blömeke, S., & Zlatkin-Troitschanskaia, O. (2013). Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor: Ziele, theoretischer Rahmen, Design und Herausforderungen des BMBF-Forschungsprogramms KoKoHs (KoKoHs Working Papers, 1). Berlin & Mainz: Humboldt-Universität & Johannes Gutenberg-Universität.

Bloom, H. S. (2009). Modern Regression Discontinuity Analysis. (7.11.2013) Retrieved from http://www.mdrc.org/publication/modern-regression-discontinuity-analysis (6.11.2013).

Dimmery, D. (2012). Introducing the `rdd' package (beta). Retrieved October 01, 2013, from http://www.drewdimmery.com/introducing-the-rdd-package-beta/

Dimmery, D. (2013). rdd: Regression Discontinuity Estimation. R package version 0.54. (7.11.2013) Retrieved from http://CRAN.R-project.org/package=rdd (6.11.2013).

Fahrmeier, L., Kneib, T., & Lang, S. (2007). *Regression: Modelle, Methoden und Anwendungen*. Berlin: Springer.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Frank, K. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods and Research, 29*, 147-194.

Frank, K. (2013). *Causal Inference and Robustness Indices.* Retrieved July 18, 2013, from https://www.msu.edu/~kenfrank/research.htm

Frank, K. A., Sykes, G., Anagnostopoulos, D., Cannata, M., Chard, L., Krause, A., & McCrory, R. (2008). Extended influence: National board certified teachers as help providers. *Education, Evaluation, and Policy Analysis, 30*(1), 3-30.

Hartmann, S. (2013). *Die Rolle von Leseverständnis und Lesegeschwindigkeit beim Zustandekommen der Leistungen in schriftlichen Tests zur Erfassung naturwissenschaftlicher Kompetenz* [The role of reading comprehension and reading speed in text-based measures of scientific inquiry skills] (Doctoral dissertation, University of Duisburg-Essen, Germany, 2013). Retrieved July 17, 2013, from http://duepublico.uni-duisburg-essen.de/servlets/DocumentServlet?id=31398

Heckman, J. (2005). The scientific model of causality. *Sociological Methodology, 35*, 1-99.

Ho, D. E., Imai, K., King, G., & Stuart, M. E. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8).

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association,* 81 (396), 945-960.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice Journal of Econometrics 142. 615–635.

Jackson, J. J., Thoemmes, F., Jonkmann, K., Lüdtke, O., & Trautwein, U. (2012). Military training and personality trait development: Does the military make the man, or does the man make the military? *Psychological Science, 23*(3), 270-277.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for english language learners in math tests. *Educational Assessment, 14*, 160–179.

National Research Council (2000). *Testing english-language learners in U.S. schools. Report and workshop summary.* Washington, DC: National Academy Press.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs* (report from the Governing Board of the American Educational Research Association Grants Program). Washington, DC: American Educational Research Association.

Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. American Psychological Association. *Psychological Methods*, 15 (1), 3-17.

Shadish, W. R. (2011). Randomized Controlled Studies and Alternative Designs in Outcome Studies: Challenges and Opportunities. *Research on Social Work Practice*, *21*(6), 636–643.

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

Shadish, W. R, Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. Psychological methods, 16(2), 179–91.

Stanat, P., & Christensen, G. (2006). *Where immigrant students succeed. A comparative review of performance and engagement in PISA 2003.* Paris: OECD.

Stuart, E. A. (2012). Software for implementing matching methods and propensity scores. Retrieved October 15, 2013, from http://www.biostat.jhsph.edu/~estuart/propensityscoresoftware.html

Teknome, K. (2007). Kernel Regression. Retrieved October 01, 2013, from http://people.revoledu.com/kardi/tutorial/Regression/KernelRegression/

Previously published:

KoKoHs Working Papers, 1

Blömeke, S. & Zlatkin-Troitschanskaia, O. (2013). Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor: Ziele, theoretischer Rahmen, Design und Herausforderungen des BMBF-Forschungsprogramms KoKoHs [Modeling and Measuring Competencies in Higher Education: Aims, theoretical framework, design, and challenges of the BMBF-funded research program KoKoHs] (KoKoHs Working Papers, 1). Berlin & Mainz: Humboldt-Universität & Johannes Gutenberg-Universität.

KoKoHs Working Papers, 2

Blömeke, S. (2013). Validierung als Aufgabe im Forschungsprogramm "Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor" [The task of validation in the research program „Modeling and Measuring Competencies in Higher Education"] (KoKoHs Working Papers, 2). Berlin & Mainz: Humboldt-Universität & Johannes Gutenberg-Universität.

KoKoHs Working Papers, 3

Blömeke, S. & Zlatkin-Troitschanskaia, O. (Eds.) (2013). The German funding initiative "Modeling and Measuring Competencies in Higher Education": 23 research projects on engineering, economics and social sciences, education and generic skills of higher education students (KoKoHs Working Papers, 3). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.

KoKoHs Working Papers, 4

Berger, S., Hammer, S., Hartmann, S., Joachim, C., & Lösch, T. (2013). Causal Inference in Educational Research. Approaches, Assumptions and Limitations. (KoKoHs Working Papers, 4). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.