Qual Quant
DOI 10.1007/s11135-009-9281-4

The use of quasi-experiments in the social sciences: a content analysis

Marie-Claire E. Aussems · Anne Boomsma · Tom A. B. Snijders

Abstract This article examines the use of various research designs in the social sciences as well as the choices that are made when a quasi-experimental design is used. A content analysis was carried out on articles published in 18 social science journals with various impact factors. The presence of quasi-experimental studies was investigated as well as choices in the design and analysis stage. It was found that quasi-experimental designs are not very often used in the inspected journals, and when they are applied they are not very well designed and analyzed. These findings suggest that the literature on how to deal with selection bias has not yet found its way to the practice of the applied researcher.

Keywords Quasi-experiments · Social science · Selection bias · Research designs · Content analysis

1 Introduction

Although randomized experiments are often seen as the golden standard for evaluating treatment effectiveness, there are many situations where the use of experimental designs is not suitable or simply impossible. In such cases, groups are compared that are often formed prior to intervention—so-called intact groups—and the treatment assignment mechanism is usually unknown. Nevertheless, the resulting quasi-experiments (cf., Shadish et al. 2002) are often analyzed using statistical methods assuming that groups were randomly composed.

M. E. Aussems (✉)
Department of Social Research Methodology, Faculty of Social Sciences,
VU University, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
e-mail: mce.aussems@fsw.vu.nl

A. Boomsma · T. A. B. Snijders
University of Groningen, Groningen, The Netherlands

T. A. B. Snijders
University of Oxford, Oxford, UK

Published online: 04 September 2009

There are countless examples of studies using quasi-experimental designs that were criticized for being not accurately enough in their investigation and conclusions. A socially relevant but controversial study in this matter is the Westinghouse report on the Head Start program (1969), developed to give poor pre-school children additional education for better first grade preparation. Another example is the Coleman Report (1966), which describes an investigation of school desegregation by comparing black children in white schools and white children in black schools. The results of these studies were doubtful because selection bias could play a major role in explaining the observed effects (cf., Campbell and Erlebacher 1970, on the Westinghouse report; Mosteller 1967; Nichols 1966, on the Coleman report).

In past decades, growing attention in social science methodology has been paid to the problems associated with quasi-experiments, especially by stressing the importance of internal validity in making proper conclusions. Internal validity refers in particular to whether an observed association found between X and Y reflects a causal relationship from X to Y in the way the variables were manipulated or measured (Shadish et al. 2002). To make the inferential step from association to causation, a researcher has to show that X precedes Y in time, that X is associated with Y, and that there are no other plausible explanations for Y than X. Especially this last condition is hard to falsify in quasi-experiments where limited knowledge of assignment mechanisms will likely lead to spurious relationships.

More efforts than before have been taken to show researchers the importance of adequately designing and analyzing research by offering guidelines (cf. Cochran 1983; Wilkinson and The Task Force on Statistical Inference 1999) and advocating a critical attitude towards their studies. Meanwhile, more advanced statistical methods were developed to deal with the possibility of selection bias (cf. Rosenbaum and Rubin 1983; Angrist et al. 1996) and overviews were written to make social scientists more familiar with alternative methods (cf. Winship and Morgan 1999; Winship and Sobel 2004). Nowadays, it is not clear to which extent these suggestions and methods are adopted by researchers and journal editors. Current practice could be observed, however, from the design and analysis of quasi-experiments that are published in social science journals.

This paper has two aims. The first is to explore how often quasi-experiments are used and how much this varies over social science disciplines. This will be investigated by identifying quasi-experimental studies that were published in selected social science journals. The second aim is to investigate how these studies deal with possible selection bias resulting from non-random assignment in the planning and analysis stage of the study. The focus will be on topics like the occurrence of inferential problems, the nature of treatments, the choice and use of covariances, the methods used to analyze the data, and the attention researchers pay to selection problems. This second goal will be accomplished by a content analysis of the quasi-experiments from the inspected social science journals.

2 Definitions

In this section a definition of a quasi-experiment is presented to distinguish it from other study designs. Furthermore, explanations are given for interventions and treatment effects to clarify the use of these terms throughout this study.

2.1 Quasi-experiments and observational studies

The designs of observational studies can roughly be divided in two groups: the analytical survey design and the quasi-experimental design (Cochran 1983). Studies with an analytical

survey design do not aim to assess the effect of an intervention on one or more responses. In using such designs, a researcher is interested in describing and predicting associations between variables, and the main objective is therefore exploration. This class of studies is often labeled correlational studies. In contrast, a quasi-experimental study is designed and conducted to investigate the effect of one variable on other ones. This design is comparable to randomized experiments in the terminology and design elements, i.e., the use of pretests and control groups. It lacks, however, one important characteristic: control with respect to the assignment of individuals to treatments. Examples of quasi-experimental designs are the nonequivalent groups design, the case control design, and the regression discontinuity design (for an overview of quasi-experimental designs, see Shadish et al. 2002).

Throughout this paper, a study is labeled quasi-experimental if two conditions hold. The first is that a researcher is interested in the effect of an intervention on one or more responses. This means that the researcher declares the *aim of the study to be causal description*, not only exploration of relationships. The investigator makes this explicit at the beginning of the study and may use language that refers to causality. The second condition is that an intervention effect is *investigated by comparing groups*, which means that control or referent groups are used for the assessment of a treatment effect. The treatment and control groups are formed or identified *before* the treatment is imposed, and it is assumed that external influences are affecting the groups to the same extent during the experimental period. Furthermore, no reference is made to the use of randomization in allocating subjects to treatment conditions.

Quasi-experimental group comparisons can be distinguished from post hoc group comparisons where groups are contrasted that were created by stratifying on specific variables *after* treatment assignment. In correlational studies the latter procedure is regularly used in correlational studies to explore how relations between variables change for subjects having different characteristics.

2.2 Interventions

There is much discussion in the literature about what should be seen as an intervention and what shouldn't (e.g., Holland 1986; Berk 2003). In a strict sense an intervention is any manipulation that can be imposed on a subject. Interventions falling into this category can be time-intensive treatments, like following a therapy or a teaching program, but also very short treatments that do not have a permanent impact on a subject, like the wording of a survey question or watching different versions of a commercial. Randomization is necessary for drawing proper conclusions: if a subject cannot be randomly assigned to any treatment it is not possible to make fair, balanced comparisons between groups receiving different treatments. Moreover, the manipulation of treatment allocation ensures that the treatment precedes the effect.

In general, however, an intervention can be seen as a much broader concept that includes treatments which are not strictly manipulable (Winship and Morgan 2007). It then mainly involves interventions that cannot be imposed on subjects for several reasons (Rosenbaum 2002): (1) The assignment of the treatment can be governed by *macro processes*. One can think of a policy rule that applies only to a selective part of the inhabitants, like financial support for educating children of parents with an income under a predefined level. This is sometimes called a natural experiment when treatment and control groups are composed in a natural way. (2) The treatment may have *harmful consequences* for a subject. It could be unethical to give some subjects education and others not, or to withhold therapy from

individuals if the therapy is known to be beneficial. (3) *Convenience*: only data from intact groups are available.

Although the class of interventions may not only consist of manipulable treatments, it is widely agreed that attributes are not included in this broader definition. Attributes are characteristics of subjects that cannot be changed by manipulation, like gender and IQ, or can only temporarily be boosted but not changed by the manipulation itself, like self-confidence and solidarity. Imposing an attribute on a subject is impossible because subjects are defined by their individual characteristics. Changing an attribute would change the subject, but the subject is assumed to be constant during the experimental period (Holland 1986). In the present study, manipulable treatments as well as non-manipulable treatments are seen as valid interventions.

2.3 Treatment effects

One important aim of quasi-experimental studies is to evaluate the change in a subjects outcome caused by receiving the treatment compared to withholding the treatment. This change can be interpreted as the *individual treatment effect* for that subject. In practice only one of these outcomes is observed, because a subject is observed in either the treatment or the control condition. To obtain the counterfactual outcome—the outcome a subject would have in the other condition—one has to find a comparable subject that received the counterfactual treatment. In practice, however, one is more often interested in the *average treatment effect* by comparing the means of a treatment and control group. The important condition for an average treatment effect to be valid is that treated and control subjects need to be similar on characteristics related to treatment assignment and the outcome variable. This latter criterion makes thoughtful design and analysis of quasi-experimental data crucial, because it is *the* way to reduce the disturbing impact of selection bias in estimating a treatment effect.

3 Method

3.1 Selection of journals

To investigate the practical use of quasi-experimental designs, a content analysis was conducted regarding 18 journals that were selected from four disciplines: psychology, criminology, education, and sociology. Journals from various social disciplines were included, because it can be expected that the use of quasi-experiments and the design and analysis of these studies varies for different journals and disciplines.

Online available journals were chosen over the period 2002–2003. Furthermore, a selection was made of journals that were more likely to include articles dealing with the evaluation of interventions. The Psycline database was searched using (combinations of) the following keywords: treatment, intervention, therapy, program evaluation, effectiveness, treatment effect, and quasi-experiment. Journals that included such words in their publisher's description were categorized regarding their discipline. For all four disciplines, journals were selected for content analysis using the following three criteria:

- *Suitability*. A journal is more suitable when the journal publisher's description contains many keywords that refer to causal inference and treatment evaluation. Articles about quasi-experiments are more likely to be found in those journals.

Table 1 Journals included for analysis

Journal	Impact factor
Psychology	
The British Journal of Social Psychology (BJSP)	1.4
Child Development (CD)	3.3
Developmental Psychology (DP)	3.4
Journal of Consulting and Clinical Psychology (JCCP)	4.2
Journal of Clinical Psychiatry (JCP)	4.8
Journal of Personality (JP)	1.9
Journal of Personality and Social Psychology (JPSP)	3.6
Personality and Social Psychology Bulletin (PSPB)	1.9
Criminology	
Criminology (C)	1.7
Criminal Justice and Behavior (CJB)	1.7
Criminal Behavior and Mental Health (CMBH)	*
(Journal of Research in) Crime and Delinquency (JCD)	1.6
Education	
Assessment and Evaluation in Higher Education (AEHE)	*
Journal of Educational Research (JER)	0.4
Journal of the Learning Sciences (JLS)	2.3
Learning and Instruction (LI)	1.6
Sociology	
American Journal of Sociology (AJS)	2.1
American Sociological Review (ASR)	2.9

* Journals having no impact factor

- *Generality*. A journal is more general when it deals with a wide range of topics within a field, instead of focusing on a specific topic only. Journals that are more general for a discipline are more appealing to a broader group of researchers.
- *Nature of research*. A journal has to include mainly quantitative research to be eligible for the present study. Although quasi-experiments can also be used in qualitative research, the latter approach is very different in the aims and reporting style than quantitative research.

After a first selection, journals with high impact factors as registered in the Social Science Citation Index 2004 were added, because those journals were underrepresented in the initial list of selected journals. In our view it was important to take this aspect into account, because it can be expected that studies published in those journals have a higher quality and are better designed and analyzed. The final list of journals is presented in Table 1.

3.2 Procedure

The journals were analyzed for the years 2002 and 2003. By trial and error it was found that searching for quasi-experiments based on keywords was not effective. Experience showed

that this procedure yields a mixture of research designs—experimental, quasi-experimental and correlational—which is a consequence of the shared use of the experimental language by quasi-experimentalists and the bad habit of large groups of correlational researchers to use terms referring to causality. Therefore, it was decided to check every article in the journals. By following such a procedure every quasi-experiment in the selected journals can, in principle, be identified.

A maximum of six issues for each journal were analyzed. When a journal published more than six issues each year, a random selection of six issues was made. The content analysis was restricted to scientific research papers and did not include editorials, comments, book reviews, etc.

All articles were analyzed independently by two encoders using a fixed procedure to examine various elements of the studies systematically. This protocol was recorded in a codebook and included descriptions of all categories to determine study characteristics. The interrater reliability was found to be good (Cohen's $\kappa = 0.78$, $p < 0.01$).

The content analysis consisted of two stages. In the first stage, a classification was made of every article to categorize the research design as an experiment, quasi-experiment, correlational study, or some other type of study. In situations where articles contain more than one study, the additional studies were analyzed as separate studies.

In the second stage, the selected quasi-experiments were analyzed in more detail. It was generally interesting to learn whether researchers take the necessary methodological steps when non-random assignment is implemented, or as Wilkinson and The Task Force on Statistical Inference (1999) suggest:"The researcher needs to attempt to determine the relevant covariates, measure them adequately and adjust for their effects either by design or by analysis" (p. 595). There were three issues under investigation:

- *Research in general*. What is the nature of the treatments, and the inference problems that are faced? What sample sizes are used, and which comparisons are made?
- *Adjustment by design*. How are the quasi-experiments designed? Which adjustments take place by design? How are covariates chosen?
- *Adjustment by analysis*. Which adjustment method is used in the analysis stage? Do researchers pay attention to limitations with respect to the internal validity of their study?

4 Results

4.1 Frequency of research designs

The 18 journals were analyzed over a total period of 2 years, which resulted in 2,618 identified studies in 2,474 articles. In Table 2 it is shown how the studies of each journal are categorized over the various designs.

The discipline making most use of the experimental design is psychology. This finding could be expected, given that randomization is easier to apply in this field where manipulations are short and researchers often face less ethical boundaries in assigning subjects to different treatments. In the journals of criminology, random assignment is less regularly used, while in the educational sciences more experiments can be found. The limited application of randomization in the latter two disciplines can partly be explained by the nature of the interventions, which are in general not fair to withhold from a subject.

The quasi-experiments make up only 4.4% of the studies. It is unknown whether this relatively low percentage reflects an unfortunate choice of journals or that the design is

Table 2 Frequencies of research designs in 18 journals

	Experimental	Quasi-experimental	(Quasi-) experimental	Correlational	Other	Total
	N	*N*	*N*	*N*	*N*	*N*
Psychology						
BJSP	20	2	12	28	17	79
CD	131	7	11	142	7	298
DP	84	11	13	101	9	218
JCCP	54	9	–	110	47	220
JCP	42	8	–	96	20	166
JP	10	–	2	92	9	113
JPSP	238	2	134	285	18	677
PSPB	85	12	30	81	4	212
Criminology						
C	2	5	–	56	13	76
CJB	5	4	–	37	22	68
CBMH	–	5	–	29	20	54
JCD	–	7	–	25	22	54
Education						
AEHE	2	8	–	34	44	88
JER	10	18	1	32	7	68
JLS	2	6	1	2	17	28
LI	22	11	4	12	22	71
Sociology						
AJS	2	1	–	27	21	51
ASR	4	–	–	49	24	77
Total	713	116	208	1,238	343	2,618

not applied very often. The 116 studies were found in 108 articles. Quasi-experiments are more often utilized in the educational sciences, somewhat less in criminology, and seldom in psychology and sociology. Most quasi-experiments (15.5%) were found in JER, while no quasi-experiments were found in JP and ASR.

A remarkable category consists of studies that fall neither clearly into the experimental group nor in the quasi-experimental group, indicated in Table 2 as '(quasi-) experimental'. These studies were not explicit about the assignment mechanisms used to construct the treatment and control groups. This lack of clarity is most noticable in psychology, where in JPSP, BJSP, and PSPB, respectively, 19.8, 15.2 and 14.2% of the studies were not explicit in this matter. A major problem of omitting such relevant information is that it is important for interpretation purposes to know whether random assignment has been applied or not. In the APA Publication Manual (2001) it is stated that: "when humans participated as the subjects of the study, report the procedures for selecting and assigning them and the agreements and payments made" (p. 18). Random assignment gives much stronger evidence for causal inference than nonrandom assignment. In the journals of criminology, sociology and the educational sciences, descriptions of assignment mechanisms were found much more explicitly.

Most studies in the four disciplines fall in the category of correlational research, i.e., studies in which researchers are mainly interested in exploration and prediction. Approximately half of all studies were correlational. The lowest proportion of correlational studies was found in JLS (7.1%).

The remaining category 'other' includes qualitative studies, overviews, and reviews of research studies. It can be seen that other research designs are more often included in criminology and the educational sciences. Especially AEHE and JLS contain many studies falling into this category (50 and 60.7%, respectively).

The general picture is that most studies are correlational. Experimental designs are most frequently used in psychology, and it is not often made very explicit that randomization was used to assign subjects to treatments. Quasi-experiments are more regularly used in JER and JLS, but in general not very often used in any of the selected journals.

4.2 General aspects of quasi-experiments

Before focussing on the specific choices in dealing with the problem of selection bias in the planning and analysis stage, some general characteristics of the identified quasi-experiments will be discussed first. It will be explored what kind of interventions were used and which inferential problems were found. Because only one quasi-experiment was found in the sociology journals, in this section no further reference will be made to this specific discipline.

4.2.1 Nature of treatment

Four different categories of treatments can be distinguished:

- The content analysis revealed that 49 studies used interventions that were in principle manipulable but the actual design lacked random assignment because of *practical limitations*. This was often the case in the educational sciences (27 studies) where groups were compared that were already following a treatment, often a teaching program. Especially in the educational sciences it is very convenient to use classes that were earlier assigned to a teaching program, rather than to assign students to different methods, which is often hard to implement. In such situations, the convenience of using intact groups has an advantage: a researcher avoids interference between children following different treatments.
- Studies in which assignment was controlled, but where not all subjects complied to the assigned treatment, can be compared to immediate drop-out in experimental designs. This results in *no full control*. Table 3 shows that there were 28 quasi-experimental studies which could be placed in this group; most of them were found in the developmental psychology journals DP and CD.
- Another category of interventions includes those that are determined by macro processes or treatments that cannot be imposed by randomization for *ethical reasons*. This type of treatments were more regularly found in the educational sciences and criminology and included interventions like the presence of a casino in town, gang membership, parental consent, and marijuana use. No studies in the social psychology journals contained such treatments.
- There was one combined category labelled '*Not manipulable/Attribute*'. There is an arbitrary boundary between the categories 'human manipulable (convenience)' and 'not manipulable'. An example is the situation where one student receives training and the other not. In some cases the teaching programs will be equivalent in content but

different in their didactical approach. The absence of randomization is then motivated by arguments of convenience. However, in cases where teaching programs are obviously different in content, it would be unethical to withhold some students a beneficial treatment. Examples of attributes are the boosting of self-esteem, degree of attitudinal ambivalence, and need for cognitive closure. These attributes were sometimes manipulated or measured by a pretest and then dichotomized by a median split. Especially PSPB contained studies using an attribute as intervention.

It can be concluded that randomization is often not applied because of convenience; in many cases in criminology it is hard to use various sentencing policies within one prison, and in the educational sciences it may not be feasible to assign class rooms randomly to different treatments. As expected, a substantial part of researchers is unable to use randomization; they will often use natural experiments in which assignment is determined by often unknown processes. The use of attributes as interventions is mainly used in social psychology.

4.2.2 Inference problem

To learn more about the way researchers tackle inference problems, it may be useful to know how limitations of the quasi-experimental design manifest themselves. As shown in Table 3, 70 quasi-experimental studies use intact groups, which is 60.3% of all inference problems that were found. In each journal the use of intact groups is the most occurring inference problem, but they occur more often in the journals on criminology.

The category non-random assignment includes studies in which assignment of subjects to groups lacks control. Non-random assignment plays a large role, especially in the educational journals (58.3%). Self-selection into treatment, drop-out, and nonresponse are less often reasons for inference problems: only 16 studies fell into this category. Self-selection is an inference problem resulting from a passive assignment mechanism used by the researcher, which will typically lead to subjects selecting themselves into a treatment they perceive to be most beneficial. The category non-response can be seen as a special case of intact groups, but it is distinguished here because these groups are passively formed. Non-response occurs whenever subjects are not willing to participate in the study or cannot be contacted for whatever reason.

It is clear from these findings that intact groups are the main cause for the occurrence of inference problems. This could be expected given the finding of the previous section that most interventions were not manipulated because of convenience. The use of unmanipulated interventions forces a researcher to use groups that were formed before treatment assignment. Other problems like self-selection, drop-out and non-response were less often found to be obstacles for making proper causal inferences.

4.3 Design elements

The use of strong quasi-experimental designs that are comparable to controlled experiments is necessary for making plausible causal inferences. To exclude possible threats to internal validity caused by the absence of randomization, some design aspects need firmer emphasis. In comparison to controlled experiments, a researcher using a quasi-experimental design needs to think harder about the assignment mechanisms of his study. This, often partly unknown, prerequisite needs to be used in the design of the study to make stronger claims of causality. This was already recognized by R.A. Fisher (Cochran 1965) when he recommended to make theories elaborate in the context of making steps from association to causation. In addition, it

Table 3 Aspects of analyzed quasi-experimental studies

Aspects	Psychology								Criminology				Education				Sociology		
	PSPB	BJSP	JPSP	P	JCP	JCCP	DP	CD	CBMH	JCD	CJB	C	AEHE	JER	JLS	LI	ASR	AJS	Total
4.2 General aspects																			
4.2.1 Nature of treatment																			
Manipulable (practical lim)	1	1	–	–	2	5	–	1	2	2	4	3	7	10	4	6	–	1	49
Manipulated (no control)	1	–	1	–	3	1	6	5	–	2	–	–	–	2	2	5	–	–	28
Manipulable (ethical lim)	–	–	1	–	3	–	5	1	1	3	–	2	1	5	–	–	–	–	22
Not manipulable/attribute	10	1	–	–	–	–	–	–	1	–	–	–	–	1	–	–	–	–	13
Other	–	–	–	–	–	3	–	–	1	–	–	–	–	–	–	–	–	–	4
4.2.2 Inference problem																			
Intact groups	10	1	2	–	–	2	9	7	4	5	4	2	3	10	4	6	–	1	70
Non-random assignment	1	–	–	–	5	2	–	–	1	–	1	–	1	6	2	5	–	–	24
Self-selection	–	–	–	–	3	–	2	–	–	–	–	1	2	1	–	–	–	–	9
Drop-out	–	–	–	–	–	1	–	–	–	–	–	–	–	1	–	–	–	–	2
Non-response	1	1	–	–	–	–	–	–	–	1	–	–	2	–	–	–	–	–	5
Other	–	–	–	–	–	4	–	–	–	1	–	2	–	–	–	–	–	–	7
4.3 Design elements																			
4.3.1 Comparison																			
1 treatment,1 control	2	–	2	–	2	2	4	3	2	5	2	3	1	10	–	4	–	–	42
1 treatment, 1 treatment	5	1	–	–	4	2	1	–	1	1	1	1	3	1	3	2	–	–	26
>1 treatment, no control	4	1	–	–	–	–	–	2	–	–	1	–	3	3	–	2	–	–	16
>1 treatment, 1 control	1	–	–	–	–	3	2	1	1	1	–	–	–	2	2	1	–	–	14
Doses treatment, 1 control	–	–	–	–	–	1	3	–	–	–	–	–	–	–	1	1	–	1	7
>1 treatment, >1 control	–	–	–	–	–	1	–	1	–	–	–	–	1	–	–	1	–	–	4
Other	–	–	–	–	2	–	1	–	1	–	–	1	–	2	–	–	–	–	7

Table 3 continued

Aspects	Psychology								Criminology				Education				Sociology		
	PSPB	BJSP	JPSP	P	JCP	JCCP	DP	CD	CBMH	JCD	CJB	C	AEHE	JER	JLS	LI	ASR	AJS	Total
4.3.2 Control group variation																			
Yes	–	–	–	–	1	–	–	–	–	–	–	1	–	–	–	1	–	–	3
No	12	2	2	–	7	9	11	7	5	7	4	4	8	18	6	10	–	1	113
4.3.3 Pretest–posttest																			
On same variables	1	–	1	–	8	9	–	1	2	–	2	2	3	4	6	8	–	–	47
On different variables	8	2	1	–	–	–	9	3	–	4	2	2	–	5	–	2	–	1	39
Posttest only	3	–	–	–	–	–	2	3	3	3	–	1	5	9	–	1	–	–	30
4.3.4 Number of pretests																			
1	9	2	2	–	8	9	9	4	2	4	4	4	2	9	4	10	–	1	83
2–4	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2	–	–	–	2
Not applicable	3	–	–	–	–	–	2	3	3	3	–	1	6	9	–	1	–	–	31
4.3.4 Number of posttests																			
1	12	1	2	–	1	2	9	4	4	6	3	2	5	15	5	5	–	1	77
2–4	–	1	–	–	4	6	1	2	1	1	1	–	2	3	1	6	–	–	29
5–7	–	–	–	–	3	1	1	–	–	–	–	–	–	–	–	–	–	–	5
>7	–	–	–	–	–	–	–	–	–	–	–	2	–	–	–	–	–	–	2
Not applicable	–	–	–	–	–	–	–	1	–	–	–	1	1	–	–	–	–	–	3
4.3.5 Choice of covariates																			
Theory	2	–	–	–	2	5	1	2	–	2	3	3	2	11	–	5	–	1	39
Explorative	–	1	–	–	1	–	–	–	–	–	–	–	–	–	–	–	–	–	2
Background variables	2	–	1	–	1	1	2	–	–	1	–	–	–	4	–	–	–	–	12
Theory + background var.	–	1	–	–	2	3	3	–	–	3	–	–	–	2	2	2	–	–	18
Unknown	8	–	–	–	2	–	5	5	2	–	1	2	2	1	–	2	–	–	30

Table 3 continued

Aspects	Psychology								Criminology				Education				Sociology		
	PSPB	BJSP	JPSP	P	JCP	JCCP	DP	CD	CBMH	JCD	CJB	C	AEHE	JER	JLS	LI	ASR	AJS	Total
Not applicable	–	–	1	–	–	–	–	–	1	1	–	–	4	4	4	2	–	–	17
4.3.6 Number of covariates																			
1–5	4	1	–	–	2	8	5	1	2	–	2	1	2	10	2	7	–	–	47
6–10	–	1	–	–	2	–	1	4	–	5	2	3	–	4	–	1	–	–	23
>10	–	–	–	–	2	1	–	1	–	1	–	–	–	1	–	–	–	1	7
Unknown	–	–	1	–	2	–	5	1	2	1	–	1	2	1	–	1	–	–	17
Not applicable	8	–	1	–	–	–	–	–	1	–	–	–	4	4	4	2	–	–	24
4.3.7 Adjustment in design stage																			
No	11	2	1	–	4	1	6	6	5	6	3	3	7	16	4	6	–	–	81
Matching prior to intervention	–	–	1	–	2	5	5	1	–	1	1	–	1	–	–	4	–	1	22
Stratification prior to intervention	1	–	–	–	–	–	–	–	–	–	–	–	–	1	–	–	–	–	2
Case-crossover	–	–	–	–	–	–	–	–	–	–	–	–	–	1	–	–	–	–	1
Historical controls	–	–	–	–	1	1	–	–	–	–	–	–	–	–	–	–	–	–	2
Other	–	–	–	–	1	2	–	–	–	–	–	2	–	–	2	1	–	–	8
4.3.8 Sample sizes																			
1–50	2	1	1	–	3	1	4	–	3	1	–	1	–	2	2	4	–	–	25
51–100	5	–	–	–	2	–	3	1	1	–	1	1	1	6	1	2	–	–	24
101–200	2	–	1	–	3	1	1	1	–	2	2	–	1	2	1	1	–	–	18
201–500	3	1	–	–	–	4	1	–	1	2	1	–	1	4	1	4	–	–	23
501–1,000	–	–	–	–	–	1	–	2	–	1	–	2	2	1	–	–	–	–	9
>1,000	–	–	–	–	–	1	1	3	–	1	–	1	1	3	1	–	–	1	13
Not applicable	–	–	–	–	–	1	1	–	–	–	–	–	2	–	–	–	–	–	4

Table 3 continued

Aspects	Psychology								Criminology				Education				Sociology		
	PSPB	BJSP	JPSP	P	JCP	JCCP	DP	CD	CBMH	JCD	CJB	C	AEHE	JER	JLS	LI	ASR	AJS	Total
4.4 Analysis stage																			
4.4.1 Adjustment in the analysis stage																			
No	12	2	2	–	6	9	10	7	5	7	4	4	7	15	6	10	–	–	106
Matching	–	–	–	–	2	–	1	–	–	–	–	–	1	3	–	–	–	1	8
New study	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1	–	–	1
Other	–	–	–	–	–	–	–	–	–	–	–	1	–	–	–	–	–	–	1
4.4.2 Model-based adjustment																			
No	9	1	2	–	5	2	6	1	4	3	1	–	6	10	5	5	–	–	60
Covariance adjustment	–	–	–	–	1	2	2	1	–	–	–	–	–	3	–	2	–	–	11
Controlling in regression	3	1	–	–	2	3	2	5	–	4	3	2	1	4	–	4	–	–	34
Propensity scores	–	–	–	–	–	–	1	–	–	–	–	–	–	–	–	–	–	1	2
Other	–	–	–	–	–	2	–	–	1	–	–	3	1	1	1	–	–	–	9
4.4.3 Modeling assumptions																			
No	12	2	2	–	1	5	9	7	2	7	4	1	5	15	6	9	–	–	87
Credibility	–	–	–	–	5	4	2	–	1	–	–	3	–	1	–	–	–	–	14
Investigates credibility	–	–	–	–	1	–	–	–	–	–	–	1	–	1	–	2	–	–	5
Consequences of violations	–	–	–	–	–	–	–	–	–	–	–	–	–	1	–	–	–	1	2
Adjustment method chosen	–	–	–	–	1	–	–	–	–	–	–	–	2	–	–	–	–	–	3
Not applicable	–	–	–	–	–	–	–	–	2	–	–	–	1	–	–	–	–	–	3
4.4.4 Discussion of limitations																			
No	10	1	2	–	3	1	8	4	3	2	–	2	5	16	3	5	–	–	65
Selection	1	–	–	–	5	1	3	2	2	4	–	1	–	2	–	–	–	1	22

Table 3 continued

Aspects	Psychology								Criminology				Education				Sociology		
	PSPB	BJSP	JPSP	P	JCP	JCCP	DP	CD	CBMH	JCD	CJB	C	AEHE	JER	JLS	LI	ASR	AJS	Total
Attrition	–	–	–	–	–	2	–	–	–	1	–	–	1	–	–	–	–	–	4
Interactions with selection	1	–	–	–	–	–	–	–	–	–	–	–	–	–	2	–	–	–	3
Maturation	–	–	–	–	–	1	–	–	–	–	–	1	1	3	1	–	–	–	7
Testing	–	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1
Instrumentation	–	–	–	–	–	–	–	–	–	–	–	–	1	–	–	–	–	–	1
Other	–	–	–	–	–	4	–	1	–	–	4	1	–	–	–	6	–	–	16
4.4.4 Referring to alternative methods																			
No	12	2	2	–	8	9	10	7	5	7	4	5	8	18	6	11	–	–	114
Not applicable	–	–	–	–	–	–	1	–	–	–	–	–	–	–	–	–	–	1	2

was Holland (1989) who claimed that causal conclusions require more observed variables in quasi-experiments than in randomized experiments in order to deal with potential selection bias. Shadish et al. (2002) recognized that adding design elements is *the* way to gather more relevant data for improving causal inference.

In this section, eight design elements (Shadish et al. 2002) are discussed: (1) The use of control groups can eliminate external effect that are affecting the treatment groups in similar ways. (2) Variation in control groups. Using control groups that differ on unobserved characteristics can give information on how hidden bias may affect causal inference. (3) The use of pretests and posttests can be an important tool to reduce selection bias. When variables are observed in a pretest, one can make treatment groups more similar before the intervention is assigned. The use of a posttest which includes measures for the same variables as in the pretest can be even more effective. (4) Number of pretests and posttests. Adding more than one pre- and posttest to a design can give information on how the outcome variable (or other variables) change over time. (5) The choice of variables to be measured are crucial in designs where randomization is infeasible. To attack selection bias in quasi-experiments it is important to consider variables that may affect the treatment assignment, so one can use these variables as covariates in adjustments before the intervention is assigned or later on in the analysis stage. (6) The number of covariates to adjust for differences between groups in the planning stage of the research. Generally, many (unknown) variables can play a role in the complex selection mechanism that affects who receives a treatment and who not. Adjusting the estimated treatment effect for one or two variables will usually not be a very prosperous attempt to reduce selection bias substantively. (7) A possible way to reduce selection bias is to make treatment groups more similar in the planning stage of the research. When a researcher identifies important covariates, it may be useful to match or stratify subjects based on variables that predict treatment assignment. Subjects are then compared to similar others which might reduce selection bias. (8) The sample sizes are discussed. One problem in making inferences is small sample size, frequently occurring in the social sciences. A large sample size is important when controlling for potential confounding variables in a statistical model, because it leads to higher power to detect a treatment effect.

4.3.1 Comparison

Randomized experiments provide information on the outcomes in the treatment and control condition. Furthermore, they reduce all kinds of biases which influence both treatment and control groups and aim therefore at uncovering the 'true' treatment effect. In experimental language often the term 'control group' is used to denote such a comparison group, although some would prefer the term 'contrast group' (cf. Wilkinson and The Task Force on Statistical Inference 1999).

There are two decisions to be made regarding the use of contrast groups. (1) A researcher must decide on the comparison to be made. The use of a comparison group implies a dichotomous comparison between a treatment group and a control group, or contrasting two different treatment groups either with or without an additional control group. To show that a treatment improves on an older one, it is most informative to contrast a new intervention with an existing one, if it is known that a comparison to a no-treatment situation is less realistic and not clearly interpretable. In the educational sciences and criminology such comparisons make sense when newly developed interventions like teaching and sentencing programs are evaluated relatively to the regularly used treatments. (2) A researcher has to choose the number of comparison groups. When a researcher decides to use multiple comparison groups, in principle more precise and powerful estimates of treatment effects can be obtained. Contrasting

a treatment group with multiple control groups creates opportunities for causal inference by getting insight in the magnitude of hidden bias, resulting from non-observed confounders that may be present in the data (Rosenbaum 2002): "Generally, the goal is to select contrast groups so that, in the observed data, an actual treatment effect would have an appearance that differed from the most plausible hidden bias" (p. 254). The aim of using multiple comparison groups is, therefore, to make a distinction between the treatment effect and systematic biases and to place bounds on the magnitude of the estimated treatment impact within known biases (cf., Campbell 1969, for a discussion on systematic variation controls and bracketing controls).

It is shown in Table 3 that half of the studies make a comparison between two groups (68 out of 116 studies). More than half of these dichotomous comparisons concern a treatment–control (T–C) situation, while the remaining groups consist of comparing two treatments (T–T). The next largest category contains articles that make a comparison between more than two treatments and include one or no contrast group. The use of different treatment groups in combination with more than one contrast group is only rarely found and the same holds for the use of multiple control groups. Only four studies identified in JCCP, DP, AEHE, and LI use more than one control group.

4.3.2 Control group variation

Varying control groups on some variable can give information about present hidden biases. However, only three studies in JCP, C, and LI use such a powerful design. Although not applicable in some studies, it seems that researchers do not fully consider the recommendations of Wilkinson and The Task Force on Statistical Inference (1999).

An unexpected finding was that relatively more interventions in the educational sciences, criminology, and psychology are contrasted to a control group than to another (traditional) treatment. Given that many studies in these disciplines are concerned with the evaluation of programs, it would enhance the interpretation of the results if the improvements were shown relatively to alternative programs. Clearly, the use of multiple control groups is not often used to attack possible threats to internal validity. Unfortunately, many researchers do not seem to recognize this approach as an effective strategy to access possible biases and the magnitude of uncertainty which results from it.

4.3.3 Pretest and posttest

In order to make adjustments in the analysis stage of the study, it is necessary to measure characteristics of the subjects *before* treatment assignment. This time point of measuring is important: the characteristics cannot be affected by the treatment. Adjusting for covariates which changed because of the intervention may remove or enhance an existing treatment effect (Rosenbaum 2002). Therefore, pretests are useful to learn more about attitudes, behavior and other characteristics that may change by the implementation of the intervention. A pretest includes variables which are either related to treatment assignment or are strong predictors of the response variables. The inclusion of measures of the response variable in the pretest is a powerful approach for removing any selection bias to be present. Correcting for the pretest value of the response variable implies that treatment groups are made similar on all variables relevant for both response and treatment assignment. Adding multiple pretests and posttests can improve the internal validity but is sometimes infeasible in applied research.

Although adding a pretest is strongly recommended, there are three reasons for researchers to use a posttest only (Shadish et al. 2002): (1) They expect that selection bias results only

from background variables that can be assessed as well after the intervention has been implemented. (2) They face a natural experiment, which refers to the situation where the intervention occurred unexpectedly and could therefore not be planned between two measurements. Retrospective pretests where subjects are asked to remember their values on some variables before they obtained treatment are sometimes used in such a situation. Unfortunately, this technique is very sensitive to recall error and leads to underestimating the treatment effect. (3) They fear that conducting a pretest may affect the responses on the posttest. Such fears are unfounded, when this systematic effect occurs in every treatment group. Usually, however, such experimentation effects—the Hawthorne effect can be seen as another example—lead to biased responses in all compared groups, but still gives unbiased estimates of the treatment effect.

It is shown in Table 3 that 47 (40.5%) of the analyzed studies used a pretest of the outcome variable. The remaining studies used a pretest without a measure for the outcome variable (39 studies) or a posttest only (30 studies). More specifically, it can be noticed that in the educational journals and the clinical psychiatry journals more often a pretest on the outcome variables is included.

4.3.4 Number of pretests and posttests

Almost 72% of the quasi-experiments used one pretest and 66.3% of the studies included one posttest. In all journals the use of one pre- and posttest was most often found, except for LI and JCCP, which included more frequently 2–4 posttests.

Although a substantial part of the studies include a pretest in their design, still almost half of those do not include pretest measures for the outcome variable. One reason for this finding could be that researchers fear learning effects when asking the same question twice. The use of posttests only is less often found in social psychology, and more regularly in the educational sciences.

4.3.5 Choice of covariates

When a researcher plans an observational study, it is important to specify the mechanisms and corresponding variables that determine treatment selection. These explanatory variables are the most suitable candidates for inclusion in a pretest. Wilkinson and The Task Force on Statistical Inference (1999) stress the need for motivating the choice of covariates and emphasize that "researchers using non-randomized designs have an extra obligation to explain the logic behind covariates included in their design and to alert the reader to plausible rival hypotheses that explain their results" (p. 600). Stated differently: in principle, the investigator has to falsify plausible alternative causes for the observed treatment effect.

In general, researchers include automatically background variables as controls in their analysis, most often without giving a theoretical motivation. In many studies these are the only variables that are included and no further effort has been taken to identify variables related to treatment assignment. In the content analysis, a distinction was made between studies including background variables, theoretical variables, a combination of both, and variables identified by explorative analysis. Although variables like gender, socio-economic status, and age can be classified as 'theoretical' or 'background', they were classified as either one of both types based on the published explanation for their inclusion.

The general picture from the content analysis is that most researchers use theoretical variables or a combination of background and theoretical measures for their study. Especially in

JER, two-thirds of the studies take both kinds of variables into consideration when performing a pretest. The remaining studies are evenly balanced as to the inclusion of theoretical variables only and background variables only. A further division by discipline does not make sense because of the low totals for the different journals. In almost one out of four studies it is unknown how covariates are selected.

The results further reveal that the reasons for the inclusion of a variable in a pretest are, in most studies, theoretical expectations of confounding and the relevance of background variables. This implies that researchers do consider possible alternative hypotheses that may lead to finding a treatment effect. The inclusion of background variables seems to be a standard routine, but is not always motivated as theoretically relevant.

4.3.6 Number of covariates

The majority of studies add one to five covariates to the analysis (47 studies). Almost 21% of the studies do not include covariates, which may lead to very biased effect estimates. Another 15% of the quasi-experiments is not explicit about the number and type of covariates that were used. This latter is very problematic for a proper judgment as to the validity of the conclusions of a study.

4.3.7 Adjustment in the design stage

When random assignment may not be feasible, it is sometimes possible to make treatments groups more comparable before treatment assignment. If researchers are aware of potentially confounding variables, they can probably match on these variables or define strata based on them. Although not every confounder may be identified and used to define strata, it is a necessary first step towards more solid causal inference. It is shown in Table 3 that only 24 quasi-experiments used matching or stratification and these studies were equally spread over disciplines. Especially in LI, JCCP, and DP, studies with adjustments in the design stage could be identified. Variables measured in a pretest are thus not always used to make treatment and contrast groups more similar *before* the intervention has been applied.

Although many researchers measure theoretically important variables and background characteristics, there is still a large proportion of studies where no covariates are used to adjust for group differences in the design stage.

4.3.8 Sample sizes

The content analysis revealed that 42% of the studies used sample sizes less than 100. Especially in social psychology journals, lower numbers of subjects were available for analysis. Large samples sizes of over five hundred were mainly found in the educational and developmental psychology journals.

4.4 Analysis stage

When researchers have adequately dealt with the problem of selection bias in the design of the study, this will be beneficial when the data have to be analyzed. The measurement of relevant variables in earlier stages of the study are important for the reduction of overt bias, which is bias that can totally be explained by differences in the values of measured covariates in the treatment and control group. Overt biases can be reduced by using covariates as

tools before statistical models are applied. As an example, it is possible to match or stratify subjects on variables strongly related to treatment assignment. In this way, treatment groups are formed that are similar and therefore comparable, and it may then be perfectly valid to perform a Student *t* test to estimate the treatment effect. The use of matching and stratification is particularly useful when (1) a researcher is uncertain whether the functional form between the covariates and treatment assignment is linear and therefore prefers a nonparametrical approach, and (2) there are a limited number of variables that predict treatment assignment. Another approach for reducing selection bias is to use analysis of covariance (ANCOVA). Covariance adjustment holds the pretest scores constant, as it is called, and examines whether there is a significant difference between the posttest scores of the treatment groups.

Alternative techniques for reducing selection bias in quasi-experimental studies are the propensity score methodology (Rosenbaum and Rubin 1983, 1984), selection models (Heckman 1979; Winship and Mare 1992), and instrumental variables (Angrist et al. 1996). These methods explicitly model—each in a slightly different way—the variables that are predictive for treatment assignment. They are making different assumptions about the selection process.

First, it was investigated how researchers analyze their quasi-experiment: do they use matching, stratification or covariance adjustment, or are they acting as if their data were obtained from a randomized experiment? Do the authors consider the assumptions that come along with such methods? Are they using alternative methods like propensity scores, selection models or instrumental variables to analyze their data? Second, it was investigated whether researchers recognize the threats in using quasi-experiments and whether they pay attention to the limitations of their own study in this respect. Do they report on potential inference problems that may be caused by selection bias?

4.4.1 Adjustment in the analysis stage

How often do researchers use matching and stratification for bias reduction? It is shown in Table 3 that matching and stratification are seldomly applied: only eight studies use matching for bias reduction. More specifically, in the social psychology journals these methods were not applied at all, while in clinical psychiatry and the educational journals they were sometimes included. The general picture is that matching and stratification are not warmly greeted by researchers in the four different fields.

4.4.2 Model-based adjustment

In most studies no adjustment at all took place in the analysis stage. Only 11 studies used covariance adjustment, and they were evenly spread over the disciplines. Given that covariance adjustment is fairly well-known in psychology, it would have been expected to find more applications of the technique there. Propensity score analysis was used in only two studies.

It can be concluded that matching, stratification and covariance adjustment (ANCOVA) are little used. Applying a Student *t* test or an ANOVA to compare means can hardly be justified in many situations. Imbalances in treatment groups can violate the assumption that both groups are random samples. If adjustment takes place in these studies, it should be done in the statistical models that are used to analyze the quasi-experimental data. Although most researchers are more inclined to adjust for pre-treatment differences in statistical models, unfortunately very few researchers recognize the advantages of using methods like matching and stratification as a first step in the analysis stage and apply them in their studies.

Researchers prefer linear regression, but hardly use analysis of covariance to adjust for mean group differences. The use of covariance adjustment is more appropriate for this aim, but the technique is hardly used in the 18 journals. This finding gives a rather pessimistic view on the attempts of researchers to attack the selection bias that may have lurked into their studies.

4.4.3 Modeling assumptions

With respect to modeling assumptions, it is observed that 75% of the studies do not reflect on the plausibility of modeling assumptions that apply. This is rather unexpected, given the fact that most studies did not perform adjustment for pre-existing differences between control groups and thus implicitly assume random assignment. Only 14 studies discussed the plausibility of the modeling assumptions, and in 7 studies the plausibility was investigated as well as possible consequences on results.

4.4.4 Discussion of limitations and alternative methods

In any study, researchers should be critical when analyzing their results and careful in making causal conclusions. This was very well recognized by Cochran (1965) who argued: "when summarizing the results of a study that shows an association consistent with the causal hypothesis, the investigator should always list and discuss all alternative explanations of his results (including different hypotheses and biases in the results) that occur to him. This advice may sound trite, but in practice is often neglected" (pp. 252–253). Similarly, a researcher should consider the assumptions of the statistical methods that were used, or as Wilkinson and The Task Force on Statistical Inference (1999) wrote:"If the effect of covariates are adjusted by analysis, the strong assumptions that are made must be explicitly stated and, to the extent possible, tested and justified" (p. 595).

The alternative explanations for the treatment effects that may be identified do not only include unmeasured variables that may have confounded the association between treatment and response. In the content analysis, some general and well-known threats to internal validity were investigated, as identified by Shadish et al. (2002):

- *Selection*. The mechanisms that assign subjects to treatments may not be random.
- *Attrition*. Drop-out from the quasi-experiment after the subjects are assigned to an intervention.
- *Maturation*. A change in value on the response variable after the intervention is imposed does not have to reflect a treatment effect, but can also be a natural development caused by maturation.
- *Instrumentation*. The measurement instruments may differ on time points of assessment or for treatment groups.
- *Testing*. The exposure to a test may influence the scores obtained on the similar test later in time. This learning effect can mistakenly be interpreted as a treatment effect.

It is shown in Table 3 that in 65 studies no attention was paid to any of such possible threats that may apply to the studies. In a small group of 26 studies selection and attrition were discussed as possible dangers for the inferences made. In three studies it was reported that some unknown variables may have moderated the treatment effect, a phenomenon also known as *treatment heterogeneity*, and in one study it was feared that dependency of observations could have disturbed the inferences. Among all studies, nine did include maturation of a subject, instrumentation, or testing as a potential threat to the internal validity. Strikingly, researchers pay hardly any attention on how to reduce such threats, for instance by using alternative

methods like the propensity score methodology, selection models and the use of instrumental variables.

Most authors did not mention possible threats to the internal validity in the discussion of their studies which may reflect a rather uncritical attitude towards the results of their study. It seems that too little effort is taken to think hard about the plausibility of the results. Making results plausible requires working systematically, and ruling out the consequences of the estimated effects if they were to be true. Where confounding by selection bias seemed a more obvious alternative explanation for the estimated effect sizes, only a small number of studies contained remarks on this issue in rounding up their conclusions.

5 Conclusions

In this article, a content analysis was performed to review the way researchers design and analyze their quasi-experimental studies. Especially in studies where randomization may be unfeasible, one should think about sophisticated designs to clarify and decrease remaining selection biases that may have sneaked inside.

What becomes clear from the results is that there is still much that researchers can learn about dealing with inference problems in the design of quasi-experiments. This conclusion fits closely a remark of Shadish et al. (2002): "However, most quasi-experiments have used very few of the potentially available quasi-experimental design elements; and our impression is that most quasi-experiments would have benefited by more attention to both the threats to inference and the design elements that might help reduce the plausibility of those threats" (p. 160). In addition, it seems fair to conclude that given that characteristics of subjects can be measured by design, researchers are not effectively using such information in the planning and analysis stage of their study. The impression is that there are no differences on this issue between journals with higher and lower impact factors. It was not expected that more sophisticated methods for causal inference in observational studies were used very often, but it is rather striking that very few researchers even adjust the means of their treatment groups for covariate imbalances.

Although an impression of empirical detail of the current practical situation on causal inference in some social science disciplines is obtained, this investigation has a few limitations. (1) The small number of studies being analyzed. The analysis of 18 journals over a time period of 2 years resulted in 116 articles using quasi-experimental designs, which is perhaps too small for making all too strong conclusions. Given that this number is spread over the different journals, one should be careful to discriminate between the different journals and disciplines. (2) The problem of generalization, i.e., the *external validity* of our conclusions. The selected journals may not be representative for the actual use of quasi-experiments in the field. Although an effort was made to select journals based on their suitability for the aim of this study, it is not certain whether the selection actually yielded the most appropriate journals. (3) The definition of the different disciplines. In selecting journals it was tried to categorize journals with respect to discipline. However, the most suitable journals for criminology were closely related to psychology. The distinction made between journals in psychology and criminology may therefore be too strict.

The importance of using strong quasi-experimental designs should never be underestimated. Wrong conclusions can push researchers into wrong directions, resulting in cumulating by misleading knowledge. Inventories of studies on specific social interventions could give an impression of the ambiguity of research findings and how different views can exist alongside for many years. Ignoring biases in observational data is not only fatal for the

validity of a study, but can eventually also have consequences for the development of substantive research in general.

Acknowledgements We would like to thank Marleen Damman for her help in encoding the articles. This study is part of the research project 'Inference for nonexperimental designs' which is supported by the Netherlands Organization of Scientific Research, Grant Number 400-03-357.

References

Angrist, J.D., Imbens, G.W., Rubin, D.B.: Identification of causal effects using instrumental variables. J. Am. Stat. Assoc. **91**, 444–455 (1996)

APA: Publication Manual of the American Psychological Association, 5th edn. APA, Washington (2001)

Berk, R.A.: Regression Analysis: A Constructive Critique. Sage, Thousand Oaks (2003)

Campbell, D.T.: Prospective: artifact and control. In: Rosenthal, R., Rosnow, R.L. (eds.) Artifact in Behavioral Research, pp. 351–382. Academic Press, New York (1969)

Campbell, D.T., Erlebacher, A.: How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In: Hellmuth, J. (ed.) Disadvantaged Child: Compensatory Education, a National Debate, vol. 3, pp. 185–210. Brunner/Mazel, New York (1970)

Cochran, W.G.: The planning of observational studies in human populations. J. R. Stat. Soc. Ser. A-G **128**, 134–155 (1965)

Cochran, W.G.: Planning and Analysis of Observational Studies. Wiley, Toronto (1983)

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., York, R.L.: Equality of Educational Opportunity. Government Printing Office, Washington (1966)

Heckman, J.J.: Sample selection bias as a specification error. Econometrica **47**, 153–161 (1979)

Holland, P.W.: Statistics and causal inference (with discussion). J. Am. Stat. Assoc. **81**, 945–960 (1986)

Holland, P.W.: Comment: it's very clear. J. Am. Stat. Assoc. **84**, 875–877 (1989)

Mosteller, F.: Preliminary report for Group D. In: Report of the Harvard Faculty Seminar on the Equal Educational Opportunity Report (1967)

Nichols, R.C.: Schools and the disadvantaged. Science **154**, 1312–1314 (1966)

Rosenbaum, P.R.: Observational Studies, 2nd edn. Springer, New York (2002)

Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **76**, 41–55 (1983)

Rosenbaum, P.R., Rubin, D.B.: Reducing bias in observational studies using subclassification on the propensity score. J. Am. Stat. Assoc. **79**, 516–524 (1984)

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

Westinghouse Learning Corporation: The Impact of Head Start: An Evaluation of the Effects of Head Start on Childrens Cognitive and Affective Development. Ohio University, Athens (1969)

Wilkinson, L., The Task Force on Statistical Inference: Guidelines and explanations: statistical methods in psychology journals. Am. Psychol. **54**, 594–604 (1999)

Winship, C., Mare, R.D.: Models for sample selection bias. Annu. Rev. Sociol. **18**, 327–350 (1992)

Winship, C., Morgan, S.L.: The estimation of causal effects from observational studies. Annu. Rev. Sociol. **25**, 659–704 (1999)

Winship, C., Morgan, S.L.: Counterfactuals for Causal Inference. Cambridge University Press, New York (2007)

Winship, C., Sobel, M.: Causal inference in sociological studies. In: Hardy, M., Bryman, A. (eds.) Handbook of Data Analysis, pp. 481–503. Sage, Thousand Oaks (2004)