DISCUSSION PAPER SERIES

IZA DP No. 270

The Evaluation of Community-Based Interventions: A Monte Carlo Study

Boris Augurzky
Christoph M. Schmidt

March 2001

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

The Evaluation of Community-Based Interventions: A Monte Carlo Study

Boris Augurzky
University of Heidelberg, and IZA, Bonn

Christoph M. Schmidt
University of Heidelberg, CEPR, London and IZA, Bonn

Discussion Paper No. 270
March 2001

IZA

P.O. Box 7240
D-53072 Bonn
Germany

Tel.: +49-228-3894-0
Fax: +49-228-3894-210
Email: iza@iza.org

IZA Discussion Paper No. 270
March 2001

ABSTRACT

The Evaluation of Community-Based Interventions: A Monte Carlo Study *

The evaluation of interventions such as active labor market policies or medical programs by means of a randomized controlled trial is often considered the gold standard. However, randomized experiments might face severe shortcomings especially if performed at the group level. One such problem is caused by small sample size which might prevent the experiment from developing its fundamental virtue in balancing all relevant covariates. This paper investigates the potential and limits of experimental and non-experimental approaches to the evaluation problem, in particular the use of instrumental variables, in a numerical simulation study, against the particular background of community-based interventions. In our simulations, we emphasize the trade-off between bias and precision by imposing a smaller number of communities whenever we model a randomized experiment, and by allowing for a correspondingly larger number of communities in all cases where selection into the program is not controlled completely by the analyst.

JEL Classification: C15, H43

Keywords: Grouped data, observational study, randomized experiment, simulation study

Christoph M. Schmidt
Department of Economics
University of Heidelberg
Grabengasse 14
69177 Heidelberg
Germany
Tel.: +49-6221-54 2955
Fax: +49 6221 54 3640
Email: schmidt@uni-hd.de

* We are grateful to the Center for Labor Economics, UC Berkeley, for its hospitality. Boris Augurzky especially acknowledges financial support by the Friedrich-Ebert-Stiftung. This research was in part supported by the Deutsche Forschungsgemeinschaft (DFG) under the research grant “Sonderforschungs-bereich (SFB) 544, Control of Tropical Infectious Diseases”.

1 Program Evaluation: The Perils of Self-Selection

Self-selection is a fundamental obstacle for the evaluation of policy interventions. In the classical case of an individually-based program, for instance a voluntary training program for unemployed workers, potential trainees will in all likelihood base their participation decision on a comparison of their perceived post-intervention outcomes with the cost of undergoing treatment. It will often be the candidates with better schooling, and the more talented or motivated individuals who tend to enter the program. As a consequence of such self-selection, analysts cannot base the assessment of program impact on a simple comparison of mean outcomes between participant and non-participant groups. Whereas it is straightforward how to tackle selection on observables – schooling in the example of the training program –, selection on unobservables – talent, motivation – provides a serious intellectual challenge.

In the overwhelming majority of applications the *mean effect of treatment on the treated*, that is the population average over the individual gains from treatment for all individuals participating in the program, is the principal object of interest. One can easily construct an estimate of the mean outcome after treatment for program participants from observed data. Yet, to perform an appropriate comparison, one has also to construct the average *counterfactual* outcome that trainees would have achieved had they not been trained, a problem of identification.

Observable data alone will not suffice to construct this entity. Researchers have proposed several alternative strategies to overcome this *identification problem*, either by invoking *a priori* information on the process of selection into treatment or other aspects of the program (Heckman & Robb, 1985, Angrist & Krueger, 1999) in a so-called observational study, or by designing an appropriate experiment. In an experiment (the classical reference is Fisher, 1935), participation is still voluntary, but some of the applicants are withheld the treatment. Who receives treatment and who does not is chosen by a random mechanism, allowing the construction of the desired counterfactual as the simple average over randomized-out controls. In the natural sciences this *randomized controlled trial* (RCT) has become the method of choice for the evaluation of interventions.

While emphasis in methodological work is on the individual, practical applications frequently concern the case of group-level or community-based interventions. Implementation of policy measures at the community-level is often a matter of necessity – whenever it would be difficult to treat some individuals in a community while excluding others, evaluation has also to be at the community-level. For instance, the evaluation of an anti-smoking information campaign at schools would require that some schools as a whole be assigned to treatment. Obviously, it would be quite cumbersome, if not impossible, to plan the intervention and its evaluation at the student level. Spill-over effects from the treated to the control students would easily contaminate the experiment. Hence, the units of interest should be groups. Moreover, analysts might choose a community-level approach to evaluation for reasons of costs. In general, interventions relevant to the social sciences often have a community-based character.

Nothing seems more natural as a methodological approach to the evaluation of community-based interventions as the translation of the RCT paradigm to the community level. Objects of randomized assignment into treatment and control samples are then entire communities, while outcomes are typically still measured at the individual level. A comprehensive overview of the theory and practice of such group-randomized trials is Murray (1998). It has long been recognized in the literature in various fields, for instance in the epidemiological literature cited in Murray (1998) and in the economics literature (see e.g. Kloek, 1981, Moulton, 1986, 1990), that the possible correlation of outcomes within communities, clusters, or groups might seriously distort conclusions regarding the statistical precision of the results.

Although one might be able to collect data on sizeable numbers of individuals within each community participating in the study, the number of communities is typically limited. If within-community correlation is substantial the effective number of observations is closely tied to the number of included communities, irrespective of the number of individuals. Thus, although group-randomized experiments implemented appropriately always produce unbiased estimates[2], it is difficult to increase precision.

Observational studies, by contrast, typically include a respectable number of communi-

[2]Contamination of randomized experiments as, for instance, attrition of treated and control units is disregarded in this paper.

ties, yet they might suffer from the selection problem. Possibly a biased but more precise estimate from an observational study might yield a lower mean squared error than the corresponding estimate of program impact from a group-randomized experiment. Thus, there might be a serious trade-off to consider in the choice of evaluation strategy (Schmidt, Baltussen & Sauerborn, 1999).

Moreover, the context of community-based interventions makes it unlikely that a randomization study can be conducted at all. Problems preventing the researcher from implementing a group-randomized experiment may be political or ethical in nature, or reflect cost considerations. However, contrary to what many practitioners apparently believe to be the state of the art – either analyze an experiment or rely on simple regression analysis to alleviate some of the disadvantages of observational data – does not properly reflect the spectrum of identification strategies for dealing with observational data. While the economic literature has long emphasized the potential of the instrumental variables method (see e.g. Bowden & Turkington, 1984, Angrist, Imbens & Rubin, 1996, Heckman, 1996), this method has not been prominent in the epidemiological literature (where it has been advocated recently by Schmidt et al., 1999).

This paper investigates the potential and limits of experimental and non-experimental approaches to the evaluation problem, in particular the use of instrumental variables, in a numerical simulation study, against the particular background of community-based interventions. In our simulations, we emphasize the trade-off between bias and precision by imposing a smaller number of communities whenever we model a randomized experiment, and by allowing for a correspondingly larger number of communities in all cases where selection into the program is not controlled completely by the analyst. We specify several variants of selection, on the individual and the community-level, and on the basis of observable and unobservable factors. Specifically, we explore the potential of instrumental variables in approximating the performance of randomized experiments (for a complementary simulation study on instrumental variables at the group-level see Shore-Sheppard, 1996).

The following section formulates the basic evaluation problem and presents several estimation techniques that have been suggested for its solution. The conceptual design

of our simulation study is explained in section 3. Section 4 discusses the results with a focus on the assessment of estimator performance, while section 5 concludes.

2 Evaluation Strategies

This section provides the formal background for our simulation study by a statement of the evaluation problem and of several solutions suggested in the literature, in particular the method of instrumental variables. Assume that Y is the outcome variable of interest. For notational convenience, subscripts indicating individuals are suppressed. Let Y_0 be the *potential* outcome if the individual would not participate in treatment and Y_1 be the *potential* outcome if the individual would. Note that only one of the potential outcomes is realized for each individual. Furthermore, let $T \in \{0, 1\}$ be a dummy variable indicating whether a unit is treated, $T = 1$, or not, $T = 0$. Under the assumption of independence of potential outcomes from the treatment status of other individuals (SUTVA, RUBIN, 1986) the expected effect of treatment on the treated unit can formally be written as

$$\Delta = \mathbb{E}(Y_1 - Y_0|T = 1). \tag{1}$$

While $\mathbb{E}(Y_1|T = 1)$ is easily identified in the subsample of all treated units, there is no way to identify the counterfactual $\mathbb{E}(Y_0|T = 1)$ unless further assumptions are imposed. The least restrictive way to gather information on $\mathbb{E}(Y_0|T = 1)$ is presented in MANSKI (1990, 1995) who demonstrates how upper and lower bounds for the counterfactual can be obtained on the basis of indisputable *a priori* information. For instance, a dichotomous outcome variable cannot take a value lower than 0 and higher than 1. In this fashion, at least some values of $\mathbb{E}(Y_0|T = 1)$ can be excluded. If one desires point estimation of the counterfactual, however, one cannot avoid either imposing additional assumptions or addressing the issue already at the stage of designing the study.

A randomized experiment is following the second route to solve this problem as follows[3]. Units who decide to participate in the program, i.e. units with $T = 1$, are randomly

[3]HECKMAN & SMITH (1995) discuss the problems of contamination that might arise in randomized experiments. Nonrandom attrition of participants or randomization biases are prominent examples. Such problems are not considered in this paper. In general, analysts stress the advantages of randomization, though. For instance, BURTLESS (1995) emphasizes the positive aspects of randomized experiments.

assigned to either an experimental *treatment group* or *control group*. The units of the control group are denied treatment and, thus, realize the potential outcome Y_0. It follows that the control group provides an unbiased estimate of the counterfactual $\mathbb{E}(Y_0|T=1)$.

Yet, while randomization at the individual level is arguably an ideal way to identify causal relationships, randomized experiments usually suffer from low sample sizes if pursued at the group level. On the other hand, observational studies do much better with regard to the sample size but additional assumptions have to be invoked to identify the counterfactual $\mathbb{E}(Y_0|T=1)$. To this purpose, several estimators have been proposed. Some of them are presented in this section including necessary assumptions that make them valid; abstracting from finite sample variations only population moments are considered.

Cross-Section and Before-After Estimators

A first assessment of an intervention might be based on comparing treated and untreated individuals after treatment occurred. Unfortunately, the mean difference of their outcomes identifies the mean effect of treatment, equation (1), only under strong assumptions on the selection process. Formally, $\mathbb{E}(Y_0|T=0)$ must be a valid substitute for the counterfactual $\mathbb{E}(Y_0|T=1)$ which requires that treated and untreated individuals be equal with respect to characteristics that rule both the selection process and the outcome equation.

Another straightforward approach to identifying the effect of an intervention rests on the availability of data for a period t' prior to treatment. In this case, the mean outcome before treatment (at time t') is compared with the outcome after the treatment (at time t), $\mathbb{E}(Y_1^t - Y_0^{t'}|T=1)$. As above, this approach requires equally restrictive assumptions to hold; otherwise following it might cause severe biases. If external disturbances over time and beyond treatment influence the outcome variable of some units these might wrongly be attributed to the intervention, producing biased estimates. For instance, it is inappropriate to perform a before-after comparison when the economic environment is characterized by cyclical swings that typically affect individuals under study.

Difference-in-Differences Estimator

A combination of the before-after comparison and the cross-section estimator leads to the difference-in-differences approach (d-i-d). It rests on the assumption that – apart from treatment – both the treated and the untreated units experience the same time-varying shocks. Assuming the time trend in the outcome variable is the same for treated and untreated units

$$\mathbb{E}(Y_0^t - Y_0^{t'}|T=1) = \mathbb{E}(Y_0^t - Y_0^{t'}|T=0),$$

the before-after comparison of the untreated group $\mathbb{E}(Y_0^t - Y_0^{t'}|T=0)$ on average reflects exactly the bias inherent in the simple before-after comparison of the treated units. Subtracting this correction term yields

$$\Delta = \mathbb{E}(Y_1^t - Y_0^{t'}|T=1) - \mathbb{E}(Y_0^t - Y_0^{t'}|T=0).$$

In other words, d-i-d requires that the difference $(Y_0^t - Y_0^{t'})$ be mean independent of the treatment T. This is violated, e.g., if the decision to participate is determined by the individual pre-treatment outcome $Y_0^{t'}$.[4] The simulation study takes account of such a selection process when opportunity costs reflected by $Y_0^{t'}$ are involved.

Instrumental Variable Estimator

Finally, instrumental variable estimation (IV) is an evaluation strategy that enjoys considerable prominence in economics (see Angrist et al., 1996 and Heckman, 1996). It has been advocated in the epidemiological literature as a possible tool to evaluate community-based interventions by Schmidt et al. (1999). Consider initially the context of a constant-effects model and assume a variable Z exists that (i) is correlated with the endogenous treatment indicator T, but that (ii) does not have a direct influence on the outcome variable Y except through T. This variable is called an *instrument* for T. The IV technique rests on the idea that the covariance between the outcome and the instrument reflects the impact of the endogenous regressor – the parameter of interest Δ – multiplied by the covariance between the regressor and the instrument.

[4] If $Y_0^{t'}$ determines selection unobserved stochastic noise in period t' will be unevenly distributed between treated and untreated units but noise in t – if only weakly correlated with that in t' – will again be more evenly distributed. Thus, the difference $(Y_0^t - Y_0^{t'})$ will depend on T.

In case of a binary instrument, the IV estimation technique provides a consistent estimator of the mean effect of treatment on the treated, resting on the ratio[5]

$$\Delta = \frac{Cov(Y,Z)}{Cov(T,Z)} = \frac{\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0)}{\mathbb{E}(T|Z=1) - \mathbb{E}(T|Z=0)}.$$

In randomized experiments, T is its own perfect instrumental variable. Since the randomized experiment is by definition independent of the outcome, and individuals perfectly comply with their treatment assignment indicated by the dichotomous indicator Z, the correlation between T and Z is 1 in absolute value (see also Heckman, 1996). Correspondingly, an instrument Z can be interpreted as a variable that is randomly distributed across units but, in contrast to a fully randomized experiment, only imperfectly induces units to behave according to its realized value. In other words, IV estimation is a *quasi-experimental* technique. Although the IV estimator is consistent if the two principal assumptions (i) and (ii) are satisfied, it might be accompanied by large variance in finite samples, especially if the correlation between the instrument and the endogenous variable T is weak (Bound, Jaeger & Baker, 1995).

A subtle issue is added to estimation with instrumental variables if the treatment effect is heterogeneous, though, i.e. if we leave the realm of the constant-effects model. Furthermore, if selection into treatment is based on the individual effects of the treatment, IV does not identify the mean effect of treatment on the treated. Rather, the IV estimator converges to the average treatment effect for all those individuals who are induced by the instrument to enter the treatment but who would have stayed off treatment otherwise. This entity is the so-called *local average treatment effect* (LATE). Recent research typically re-interprets the IV estimate as a LATE, see e.g. Angrist et al. (1996) and Heckman (1997). Its peculiarities are discussed in section 4.

[5]The second equation is easily verified

$$\begin{aligned} Cov(Y,Z) &= \mathbb{E}(YZ) - \mathbb{E}Y\mathbb{E}Z \\ \mathbb{E}(YZ) &= \mathbb{E}(Y|Z=1)\mathbb{P}(Z=1) \\ \mathbb{E}Y &= \mathbb{E}(Y|Z=1)\mathbb{P}(Z=1) + \mathbb{E}(Y|Z=0)\mathbb{P}(Z=0) \\ \mathbb{E}Z &= \mathbb{P}(Z=1), \end{aligned}$$

Thus, $Cov(Y,Z) = \mathbb{E}(Y|Z=1)\mathbb{P}(Z=1)\underbrace{(1-\mathbb{P}(Z=1))}_{=\mathbb{P}(Z=0)} - \mathbb{E}(Y|Z=0)\mathbb{P}(Z=0)\mathbb{P}(Z=1)$. Likewise, $Cov(T,Z)$ is transformed.

Conditioning on Observables

Whenever researchers succeed in capturing *observable* elements of the process jointly determining outcomes and program participation, they can improve upon their evaluation strategy by conditioning on these observable covariates. In anticipation of the simulation setup implemented below, let X be an explanatory binary variable that takes the values 0 and 1. If self-selection depends in part on the realization of X, then within the subsamples defined by $X = 0$ and $X = 1$, any remaining bias can only reflect the presence of other factors. The selection bias would even disappear completely if selection depended exclusively on X apart from random disturbances.

Then, participation is purely *random* within the two subsamples characterized by $X = 0$ and $X = 1$, i.e. T is independent of (Y_0, Y_1) given X. It follows that the untreated units in the subsamples $\{X = x, T = 0\}$ provide the counterfactual $\mathbb{E}(Y_0|X = x, T = 1)$.[6] The unconditional mean $\mathbb{E}_X\,\mathbb{E}(Y_0|X, T = 1)$ is obtained as weighted average over the conditional means. In sum, if $\mathbb{E}(Y_0|X = x, T = 1) = \mathbb{E}(Y_0|X = x, T = 0)$ it follows

$$\begin{aligned} \mathbb{E}(Y_1 - Y_0|T = 1) &= \sum_{x\in\mathcal{X}} \mathbb{E}(Y_1 - Y_0|X = x, T = 1)\mathbb{P}(X = x) \\ &= \sum_{x\in\mathcal{X}} (\mathbb{E}(Y_1|X = x, T = 1) - \mathbb{E}(Y_0|X = x, T = 0))\mathbb{P}(X = x) \end{aligned} \tag{2}$$

where $\mathcal{X}$ is the set of all possible values of X.

Since *unobservable* variables might additionally play a role in determining the selection process, conditioning on observables alone might not enable the researcher to avoid selection bias. Yet, at least conditioning on observables might achieve to mitigate the problem. Because of that, it is recommended whenever possible. In this study, it is very easy to follow this advice due to the binary nature of the observable variables. In practice however, X might be of high dimension, thus, conditioning on all observables would be quite cumbersome or even impossible. This problem is alleviated by imposing a regression model or by conditioning on the *propensity score*, which is the probability of participation given the observable variables. Then, the sample would be stratified into subsamples of

[6] However, in the extreme case when selection is fully determined by an observable X without any stochastic components left the set $\{X = x, T = 0\}$ is either empty or equals $\{X = x\}$ making it impossible to obtain the counterfactual from untreated individuals.

units with equal or similar propensity scores and the overall treatment effect estimate would be constructed as a weighted average as in (2); the technique is often referred to as *matching*. For further discussion of this topic, see e.g. Rubin (1973, 1974), Rosenbaum & Rubin (1983, 1984, 1985), and Heckman, Ichimura & Todd (1997).

3 The Simulation Setup

The simulation is based on a data generating process that consists of two main equations, the outcome and the selection equation, and two time periods, one before and one after treatment. While the outcome equation always combines observable and unobservable characteristics with heterogeneous treatment effects, we consider two conceptually distinct modes of selection into treatment. In one set of experiments, selection into treatment is at the individual level – here we do not expect group level variables to introduce any fundamental difficulties; we also consider situations, though, in which selection into treatment is decided upon at the group level. It is these simulations where we particularly expect new insights to emerge from our simulations.

The outcome Y_{igt} of individual i, $i = 1, ..., n_g$, in group g, $g = 1, ..., G$, at time $t \in \{0, 1\}$ depends linearly on time-invariant individual and group characteristics X_{1ig} and X_{2g}, respectively, which are observable, as well as on unobservable characteristics, ν_{1ig} and ν_{2g}. Furthermore, a variable μ_t captures exogenous time-variant shocks being constant for all individuals in a given time period but displaying an upward time trend. The unobservable variables ε_{1igt} and ε_{2gt} reflect white noise at the individual and the group level. The treatment effect is a sum of an individual effect δ_{1ig} and a group effect δ_{2g} which are both random variables resulting in heterogeneity in the impact of treatment across individuals and groups. The dichotomous variable T_{igt} indicates the treatment status. In sum,

$$Y_{igt} = \alpha_0 + \alpha_1 X_{1ig} + \alpha_2 X_{2g} + (\delta_{1ig} + \delta_{2g}) T_{igt} + \nu_{1ig} + \nu_{2g} + \mu_t + \varepsilon_{1igt} + \varepsilon_{2gt}. \qquad (3)$$

In our simulations X_1 and X_2 are binary variables taking the values 0 and 1 with equal probability. Both treatment effects δ_{1ig} and δ_{2g} follow a normal distribution with mean $\frac{1}{2}$ and variance of $\frac{1}{4}$, the ν's and ε's are distributed normally with mean zero and variance

Table 1: **Variables and Parameters.**

Variable	Comment	Parameter	Comment
Outcome Equation			
Constant	–	α_0	0
X_{1ig}	binomial(1, $\frac{1}{2}$)	α_1	1
X_{2g}	binomial(1, $\frac{1}{2}$)	α_2	1
ν_{1ig}	$\mathcal{N}(0, \frac{1}{4})$	1	–
ν_{2g}	$\mathcal{N}(0, \frac{1}{4})$	1	–
δ_{1ig}	$\mathcal{N}(\frac{1}{2}, \frac{1}{4})$	1	–
δ_{2g}	$\mathcal{N}(\frac{1}{2}, \frac{1}{4})$	1	–
ε_{1igt}	$\mathcal{N}(0, \frac{1}{2})$	Corr$(\varepsilon_{1ig0}, \varepsilon_{1ig1})$	0.25
ε_{2gt}	$\mathcal{N}(0, \frac{1}{2})$	Corr$(\varepsilon_{2g0}, \varepsilon_{2g1})$	0.25
μ_t	$\mathcal{N}(\frac{1}{2}t, \frac{1}{16})$	1	–
Cost Equation			
Constant	–	τ_0	such that 50% of the sample participate
Z_{1ig}	binomial(1, $\frac{1}{2}$)	τ_1	suitable for given correlation (Z_1, T)
Z_{2g}	binomial(1, $\frac{1}{2}$)	τ_2	suitable for given correlation (Z_2, T)
η_{ig}	$\mathcal{N}(0, 1)$	1	–

The variables are independently and identically distributed if not mentioned otherwise.

$\frac{1}{4}$ and $\frac{1}{2}$, respectively. Both individual and group ε's are positively correlated over time with value 0.25, and $\mu_0 \sim \mathcal{N}(0, 1/16)$ and $\mu_1 \sim \mathcal{N}(0.5, 1/16)$. The constant α_0 equals 0 while $\alpha_1 = \alpha_2 = 1$. Table 1 summarizes parameters and variables.

When selection into treatment is considered to be an individual decision, it is modeled as an optimization process as in Heckman, LaLonde & Smith (1999: ch. 8). Individuals decide to participate if they expect to gain from treatment and, thus,

$$T_{igt} = \begin{cases} \mathbf{1}[G_{ig} > 0] & : \quad t = 1 \\ 0 & : \quad t = 0. \end{cases} \tag{4}$$

The individual net gain G_{ig} represents the difference between benefits and cost of treatment. The benefits comprise all future treatment effects $\delta_{1ig} + \delta_{2g}$ discounted to present value assuming a constant discount factor of 0.1 and constant effects beyond period $t = 1$. The cost of treatment is the sum of opportunity costs and other costs C_{ig} to be specified

below. The opportunity costs of undergoing treatment comprise outcome before treatment Y_{ig0} reflecting the presence of observable and unobservable characteristics. Finally, net gains are contaminated by stochastic noise η_{ig}.

It has been recognized in the evaluation literature (see HECKMAN, 1997) that the information available to individuals at the time of their decision whether to participate in a program is a decisive element of the selection effects to be expected. Specifically, if individuals know their own treatment effect and act upon it, the presence of heterogeneous treatment effects will necessarily lead – *ceteris paribus* – high-impact individuals to be over-represented among the individuals receiving treatment. In consequence, the mean effect of treatment on the treated will exceed the population average of the treatment effects.[7]

On the other hand, individuals acting upon the precise knowledge of their opportunity costs during the treatment period $t = 0$ will – *ceteris paribus* – typically choose to receive treatment if their time-invariant characteristics generate relatively low outcomes in both periods. While observable characteristics are controlled for easily enough, it is the unobservables which create the *selection effects* any successful evaluation strategy has to deal with. We will consider situations in which individuals select treatment on the basis of information on (i) opportunity costs Y_{ig0} and their expectation of treatment $\mathbb{E}(\delta_{1ig} + \delta_{2g})$, on (ii) precise information about both Y_{ig0} and $\delta_{1ig} + \delta_{2g}$, and on (iii) *expected* opportunity costs $\mathbb{E}Y_{ig0}$ and on expected effects $\mathbb{E}(\delta_{1ig} + \delta_{2g})$ conditional on time-invariant characteristics, respectively[8]. These various alternatives of G_{ig} can generally be written as

$$G_{ig} = \mathbb{E}\left(\frac{\delta_{1ig} + \delta_{2g}}{0.1}\middle|\,\Omega\right) - \mathbb{E}\left(Y_{ig0}|\Omega\right) - C_{ig} + \eta_{ig} \tag{5}$$

with $\mathbb{E}(\cdot|\Omega)$ denoting a conditional expectation given information set Ω. This set may contain a subset of all relevant variables, but may also contain all variables, observable and unobservable, rendering the expectation operator unnecessary. Thus, depending on the fineness of Ω either one or even both of the two expectation terms in equation (5)

[7]Naturally, as long as the evaluation strategy will be able to identify the mean treatment effect for this subpopulation, this is not a fundamental flaw of the setup, but rather a beneficial consequence of the liberation from a constant-effects model.

[8]The timing of treatment choice and outcome realization renders the scenario $\mathbb{E}Y_{ig0}$ and $\delta_{1ig} + \delta_{2g}$ irrelevant.

may coincide with identity.

Other costs C_{ig} allow the introduction of instrumental variables most naturally. Consider costs being a function of two variables Z_{1ig} and Z_{2g}, where Z_{1ig} is defined at the individual and Z_{2g} at the group level; they take the values 0 and 1 with probability 0.5 each. These variables reflect aspects such as, for example, the distance to the treatment site. In effect, other costs are

$$C_{ig} = \tau_0 - \tau_1 Z_{1ig} - \tau_2 Z_{2g}. \tag{6}$$

The constant τ_0 is chosen such that 50% of all units undergo treatment[9] and τ_1 and τ_2 are adapted such that the correlation between the instruments Z_1 and Z_2 and treatment choice T correspond to a given value (see also table 1)[10].

Treatment choice is a completely different matter if it is decided upon at the group level. Most importantly, if one of the individuals in a group receives treatment, so do all other members of the group. In our study some democratic majority decision rule is supposed to govern treatment choice. That is, the form of the selection equation (4) is retained, albeit with the group specific expected gain G_g as its argument, and thus $T_{ig1} = \mathbf{1}(G_g > 0)$. The gain G_g of a group is simply the sum over all individual expected gains. The same information scenarios arise as under individual treatment choice. Thus, groups join treatment if their aggregate expected gain is positive. Note that summing up individual gains G_{ig} of the group members reduces considerably the importance of all individual level variables as far as selection into treatment is concerned and the group variables clearly dominate the decisions.

Any observational study would proceed along the following lines: take the sample of treated and untreated individuals (with treatment varying within groups or not), and observed individual-level and group-level characteristics (X and Z), under a specific identification assumption, e.g. mean independence of treatment and outcomes conditional on observable X. Alternatively, one might be able to tackle the evaluation problem by design, namely by constructing a randomized controlled trial. In this study, we consider

[9] In fact, this is done by replacing the criterion $G_{ig} > 0$ in equation (4) by $G_{ig} > median(G_{ig})$.

[10] The costs might be dependent on the other covariates X and ν, too, but this would complicate the setup without further illuminating the main aspects of the simulation study.

randomized experiments which are performed at the individual and at the group level. Throughout, these experiments are assumed not to be contaminated by attrition or by randomization bias and, throughout, they recruit their volunteers from the pool of individuals (or groups) who are willing to participate, i.e. those with a positive net gain. Irrespective of the level of implementation, these experiments identify the mean effect of treatment on the treated under all combinations of parameters.

However, since randomized experiments, in particular those conducted on the group level, usually suffer from small sample size, the corresponding impact estimates might display a high variance compared to estimates of large scale observational studies: although an experiment achieves to balance all covariates *on average*, it might drastically fail to do so in a particular small sample. To alleviate this problem, our simulated randomized experiments follow the recommendation to stratify samples prior to randomization with respect to observable covariates and then to perform randomization within the strata (Murray, 1998). This procedure ensures that at least observable covariates are balanced.

Nevertheless, the sample size of randomized trials is comparatively small. Thus, the number of groups in randomized experiments is set equal to 20 while the corresponding number in observational studies varies between at least 40 and up to 300. This range is used to investigate the relative performance of estimates produced by observational studies and randomized experiments. In both experimental and observational scenarios, each group consists of 50 individuals. In the field, it typically is the involvement of further communities, not of individuals within communities, which raises the cost of a study.

4 Simulation Results

The discussion of results focuses on two main features. First, all estimators of section 2 are presented for different more or less favorable scenarios both at the group and at the individual level and compared with regard to the root mean squared error (RMSE). Apart from root mean squared errors – reported in squared parentheses – variation net of bias, calculated as $\sqrt{RMSE^2 - bias^2}$, will be reported in round parentheses. This serves to assess the importance of the estimator's bias component when switching from

one scenario to another. Second, separate more extensive simulations are performed to compare particularly the quasi-experimental technique of IV with fully randomized experiments.

Individual Level Selection

Table 2 is dedicated to estimation results when selection into treatment occurs at the individual level. In the basic scenario reported in column (1) only observable variables determine both the outcome and the selection equation. In particular, selection depends on the expected treatment effect $\mathbb{E}(\delta_1 + \delta_2)$, similarly, opportunity costs are captured by $\mathbb{E}(Y_0|X)$, while ν and μ are excluded from the equations. Thus, as documented in column (1), identification problems do not arise except for the simple cross-section estimator that does not control for X and thus misses to control for self-selection. The difference in RMSE between the cross-sectional and the before-after estimator is due mainly to the fact that the first is based on more observations than the latter even though correlated ε's over time help reduce the RMSE of the before-after comparison. Increasing this correlation would successively diminish the RMSE of the before-after comparison.[11] Similarly, the d-i-d estimator is affected by this correlation, too.

On the other hand, IV based on the individual instrument Z_1 suffers from the largest RMSE among all non-experimental estimators owing to its high variance but not to inconsistent estimation; a fact that is common in IV estimation: the lower the correlation between instrument and endogenous regressor T the larger the variance of the IV estimate. The coefficients of Z_1 and Z_2, τ_1 and τ_2, are chosen such that the correlations are approximately 0.3,[12] the realized values and their standard deviations across simulation iterations are shown in the table. In practice, such correlations are usually considered producing good instruments (Hall, Rudebusch & Wilcox, 1996). Further note that IV estimation controlling for X does not reduce the RMSE which can be explained by the independence of Z and X in this simulation study. Moreover, the IV estimates based on

[11] The conditional before-after-comparison is omitted since X-variables are time-constant and thus the conditional and unconditional estimates coincide.

[12] The values of τ_1 and τ_2 are adapted step by step by performing additional simulations until the correlations take the desired value.

Table 2: **Estimation Results, Selection at the Individual Level.**

Estimators	Basic (1)	ν's included (2)	μ's included (3)	Indiv. opport. costs (4)	Indiv. treatm. effects (5)
True effect	**0.999**	**0.999**	**0.999**	**1.000**	**1.466**
	(0.040)	(0.040)	(0.039)	(0.039)	(0.039)
Standard Estimators					
Cross-section	**0.430**	**0.033**	**0.032**	**0.012**	**1.233**
	(0.050)	(0.066)	(0.068)	(0.071)	(0.085)
	[0.572]	[0.969]	[0.970]	[0.990]	[0.248]
– controlled for X	**0.999**	**0.458**	**0.456**	**0.360**	**1.326**
	(0.045)	(0.062)	(0.062)	(0.066)	(0.074)
	[0.045]	[0.545]	[0.547]	[0.643]	[0.159]
Before-after	**1.000**	**1.001**	**1.495**	**1.796**	**2.030**
	(0.067)	(0.068)	(0.363)	(0.357)	(0.362)
	[0.067]	[0.068]	[0.614]	[0.872]	[0.671]
Difference-in-differences	**0.999**	**1.000**	**0.998**	**1.586**	**1.605**
	(0.054)	(0.061)	(0.061)	(0.064)	(0.074)
	[0.054]	[0.061]	[0.061]	[0.590]	[0.157]
– controlled for X	**0.999**	**1.000**	**0.998**	**1.638**	**1.605**
	(0.053)	(0.062)	(0.062)	(0.066)	(0.074)
	[0.053]	[0.062]	[0.062]	[0.642]	[0.157]
Instrumental Variables Estimators					
IV Z_1	**0.999**	**0.999**	**1.001**	**1.003**	**1.000**
	(0.087)	(0.101)	(0.103)	(0.102)	(0.102)
	[0.087]	[0.101]	[0.103]	[0.102]	[0.478]
– controlled for X	**0.997**	**1.001**	**1.001**	**1.003**	**0.996**
	(0.090)	(0.100)	(0.101)	(0.096)	(0.091)
	[0.090]	[0.100]	[0.101]	[0.096]	[0.479]
– Corr(Z_1, T)	**0.302**	**0.301**	**0.301**	**0.298**	**0.299**
	(0.010)	(0.011)	(0.011)	(0.011)	(0.013)
IV Z_2	**1.023**	**1.042**	**1.026**	**1.018**	**0.998**
	(0.432)	(0.509)	(0.510)	(0.511)	(0.506)
	[0.433]	[0.511]	[0.510]	[0.512]	[0.689]
– controlled for X	**1.003**	**1.042**	**1.022**	**1.030**	**0.980**
	(0.404)	(0.481)	(0.473)	(0.477)	(0.461)
	[0.404]	[0.483]	[0.474]	[0.478]	[0.670]
– Corr(Z_2, T)	**0.300**	**0.299**	**0.299**	**0.298**	**0.298**
	(0.022)	(0.026)	(0.026)	(0.029)	(0.036)
Experimental Evaluations					
Experiment	**0.999**	**0.996**	**0.998**	**0.998**	**1.465**
	(0.097)	(0.101)	(0.104)	(0.103)	(0.087)
	[0.097]	[0.101]	[0.104]	[0.103]	[0.087]
Stratified experiment	**0.998**	**0.996**	**0.996**	**0.997**	**1.465**
	(0.099)	(0.108)	(0.107)	(0.108)	(0.092)
	[0.099]	[0.108]	[0.107]	[0.108]	[0.092]

Means over all simulation iterations. Root mean squared errors are in square parentheses. Round parentheses show variation net of bias. Number of groups in observational studies: 200, number of iterations: 5000. Controlling for X in the before-after-comparison does not change estimation results and is omitted.

the grouped instrument Z_2 are accompanied by substantially higher variance caused by the intra-group correlation among group members which reduces the effective sample size. The detrimental effects of intra-group correlations on the precision of impact estimates is the topic of a large literature in economics (Kloek, 1981, Moulton, 1986) and epidemiology (e.g. Murray, 1998). Recently, Shore-Sheppard (1996) has extended this discussion to the problem of grouped instruments.

Experimental estimates are reported in the last two rows of the table. Taking into account their small sample size they still perform quite well: compared to the large observational studies consisting of 200 groups or 10000 individuals, the randomized experiments have to rely on only 20 groups or 1000 individuals. Yet, under these circumstances, one would prefer the standard observational approaches and the IV estimate using the individual-level instruments.

Column (2) reports estimates if unobservable characteristics ν enter the outcome equation. Naturally, this does not influence the true effect but poor performance (= low opportunity costs) might mistakenly be attributed to a poor effect of the treatment. Obviously, the cross-section estimator breaks down because it is unable to control for the unobservable ν's. However, since the ν's are time-constant, both the before-after comparison and the d-i-d are entirely unaffected by them, the unobservables just cancel out, and the estimates do not display higher variance. This is in contrast to IV: though it is consistent as a cross-sectional estimator its variance increases to a small extent. The experimental estimator, however, remains basically unaffected and performs as well as the IV estimator.

Including time-variable shocks μ into the outcome equation destroys the before-after comparison (column 3) while the other estimators remain unaffected. Since all individuals experience a higher outcome in the post-treatment compared to the pre-treatment period irrespective of having received treatment or not, the before-after comparison wrongly attributes this general increase to the treatment. D-i-d successfully achieves to correct for this bias by exploiting the before-after comparison of the untreated units who experienced the same upward trend as the treated individuals, thus serving as controls.

The fourth column shows estimates when more severe endogeneity problems are in-

troduced. Opportunity costs are assumed to be captured by individual outcomes before treatment, Y_{ig0}, instead of their conditional expectation, $\mathbb{E}(Y_{ig0}|X, \nu, \mu)$, as above. Consequently, ε_{1ig0} and ε_{2g0} determine selection, too, so they systematically differ between treated and untreated. Since the ε's vary over time and units, the d-i-d estimator cannot difference out the bias caused by them. The problem is the more severe the less is the correlation between ε's before and after the intervention. In the simulation the correlation is set equal to 0.25; a perfect correlation would eliminate the bias in the d-i-d estimate. Note that the first two non-experimental estimators are also negatively affected in this scenario where more unobservables than before rule selection. Specifically, the before-after-comparison suffers from regression to the mean: while ε_{1ig0} and ε_{2g0} determine selection and hence are unevenly distributed across treated and untreated individuals, ε_{1ig1} and ε_{2g1} are more evenly distributed depending on the strength of the correlation ρ. On the other hand, IV still produces estimates of the same quality as before, that is, IV based on Z_1 and the experimental approach are the preferred estimation strategies.

Finally, column (5) presents results of a scenario that additionally assumes that individuals correctly anticipate their individual gain $(\delta_{1ig} + \delta_{2g})$ from participation. Although the assumption is strong, it might be fulfilled if the setup of the program is transparent to the public and people are able to judge well how they succeed in the treatment. In this case, only the most successful individuals undergo treatment and, therefore, the mean effect on the treated increases from 1 to 1.47.

Under this optimization behavior IV breaks down[13]; it does not anymore identify the mean effect of treatment on the treated but the so-called *local average treatment effect* (LATE) which is the effect of treatment on someone who complies with the instrument, i.e. who participates, $T = 1$, if $Z = 1$ and who does not, $T = 0$, if $Z = 0$. In accordance with Angrist et al. (1996) and Imbens & Angrist (1994) further denote *always-takers* as individuals who always undergo treatment irrespective of the realization of their Z and, likewise, *never-takers* as those who never participate.[14] Disregarding the group

[13]Notice that the heterogeneity is not caused by observable covariates which could be coped with, but by hidden characteristics only known to the individual.

[14]A fourth category, so-called *defiers* who simply do the opposite of what their Z indicates are ruled out since T is monotonic in Z.

variables for simplicity *compliers* are characterized by the set

$$\{i : 10\delta_i - Y_{i0} - \tau_0 + \eta_i \leq 0 \quad \text{and} \quad 10\delta_i - Y_{i0} - \tau_0 + \tau_1 + \eta_i > 0\}. \tag{7}$$

Since exactly half of the sample undergoes treatment it can be shown that the individual treatment effects of never-takers, compliers, and always-takers are ordered symmetrically around the mean value of δ_i with never-takers at the bottom, compliers in the middle, and always-takers at the top. Thus, under these special circumstances, the mean effect on compliers coincides with the mean effect on a randomly chosen person, namely 1. On the other hand, the mean effect on the treated – the always-takers and compliers who participate – exceeds 1. If the selection criterion were replaced such that exactly 40% of the sample participated, LATE would increase to above 1, if 60% underwent treatment LATE would fall below 1. LATE answers the question of how large the gain from treatment would be if the costs C_{ig} were reduced by τ_1 (or τ_2 in case of Z_2).[15]

Alas, the other non-experimental estimators do improve in this scenario which is due mainly to adverse effects which cancel out some biases. No general pattern underlies this improvement. It is merely an artefact of the special model used here. In this last scenario, solely the randomized experiment is still able to identify the parameter of interest.

Interestingly, the stratified experiments in all scenarios do not do better than the unstratified ones. This is not completely unexpected, since with a sample size of 1000 randomization already balances the two-dimensional observable covariates X. This will be demonstrated to be different in the context of group randomization.

Selection at the Group Level

If treatment occurs at the group level, the effective sample size shrinks because the limited variability of treatment receipt within groups confronts similarly limited variability of observable and unobservable characteristics. In fact, results presented in table 3 clearly demonstrate how RMSE's have increased, particularly because the estimator's variances

[15]Moreover, the set of compliers (7) offers an interesting intuitive interpretation of the relationship between instrumental relevance measured by correlation between Z and T and the variance of the IV estimator. High correlation induces large τ_1 and thus a large set of compliers which, in turn, increases the number of observations IV is based on, i.e. reduces its sample variance.

Table 3: **Estimation Results, Selection at the Group Level.**

Estimators	Basic (1)	ν's included (2)	μ's included (3)	Indiv. opport. costs (4)	Indiv. treatm. effects (5)
True effect	**0.999**	**0.999**	**1.001**	**1.000**	**1.364**
	(0.050)	(0.051)	(0.051)	(0.051)	(0.043)
Standard Estimators					
Cross-section	**0.284**	**-0.057**	**-0.054**	**0.059**	**1.185**
	(0.114)	(0.122)	(0.120)	(0.126)	(0.143)
	[0.724]	[1.064]	[1.062]	[0.949]	[0.229]
– controlled for X	**1.001**	**0.287**	**0.294**	**0.358**	**1.257**
	(0.177)	(0.164)	(0.161)	(0.134)	(0.125)
	[0.177]	[0.731]	[0.725]	[0.655]	[0.165]
Before-after	**0.999**	**0.998**	**1.503**	**1.768**	**1.919**
	(0.087)	(0.087)	(0.366)	(0.363)	(0.363)
	[0.087]	[0.087]	[0.621]	[0.850]	[0.663]
Before-after, no ε_1	**0.999**	**0.997**	**1.503**	**1.874**	**1.941**
	(0.122)	(0.122)	(0.376)	(0.373)	(0.373)
	[0.122]	[0.122]	[0.627]	[0.950]	[0.687]
Difference-in-differences	**0.999**	**1.000**	**1.001**	**1.548**	**1.475**
	(0.123)	(0.122)	(0.124)	(0.117)	(0.123)
	[0.123]	[0.122]	[0.124]	[0.561]	[0.165]
– controlled for X	**1.000**	**1.000**	**1.001**	**1.648**	**1.475**
	(0.217)	(0.173)	(0.177)	(0.133)	(0.124)
	[0.217]	[0.173]	[0.177]	[0.661]	[0.167]
Instrumental Variables Estimators					
IV Z_2	**1.041**	**1.101**	**1.068**	**1.065**	**1.000**
	(0.496)	(0.730)	(0.590)	(0.649)	(0.530)
	[0.497]	[0.737]	[0.594]	[0.652]	[0.643]
– controlled for X	**1.004**	**1.089**	**1.049**	**1.104**	**1.037**
	(0.540)	(0.680)	(0.625)	(0.825)	(4.545)
	[0.540]	[0.686]	[0.627]	[0.831]	[4.557]
– Corr(Z_2, T)	**0.307**	**0.295**	**0.302**	**0.288**	**0.306**
	(0.067)	(0.067)	(0.067)	(0.067)	(0.067)
Experimental Evaluations					
Experiment	**0.998**	**1.007**	**1.006**	**1.006**	**1.362**
	(0.390)	(0.415)	(0.408)	(0.432)	(0.463)
	[0.390]	[0.416]	[0.408]	[0.432]	[0.463]
Stratified experiment	**0.996**	**0.999**	**1.006**	**0.996**	**1.363**
	(0.360)	(0.397)	(0.398)	(0.402)	(0.427)
	[0.360]	[0.397]	[0.398]	[0.402]	[0.427]

Means over all simulation iterations. Root mean squared errors are in square parentheses. Round parentheses show variation net of bias. Due to negligible correlation between Z_1 and T corresponding results are not meaningful and left out. Number of groups in observational studies: 200, number of iterations: 5000.

have done so. However, the main pattern of results remains almost unchanged compared to table 2. The RMSE's of the cross-section estimator increases whereas that of before-after comparison only slightly rises. This is because a before-after-comparison still works at the individual level since all group variables, which are time-constant, just cancel out and individual level variation caused by ε_{1igt} gains the upper hand again. Therefore, table 3 presents an additional before-after-estimator based on data where ε_1 is removed and the variance of ε_2 is increased to 1. Then, the efficiency of this estimator worsens, too.

The difference-in-differences estimator doubled its RMSE and that of its counterpart controlling for X is even three times larger. Controlling for X would reduce bias caused by X, though, it increases the variance because subsamples defined by X might be rather small, specifically at the group level. Albeit, all standard estimators continue to be consistent. Concerning instrumental variables estimation, only Z_2 is a relevant instrument while correlation between Z_1 and T is negligible and therefore results are omitted.[16] Compared to table 2 the grouped IV estimates display higher variance which might be attributed to substantially increased variance of the instrumental correlation. If in some iteration of the simulation the correlation happens to be very small, close to zero, this iteration will contribute an extremely high variance to the mean over all iterations, particularly if X is controlled for. Yet, there are no grounds for failure of the IV in identifying the treatment effect. Finally, the experimental estimator's variance quadrupled but achieves to outperform IV. IV estimation and the randomized experiment will be compared in detail under several settings below. Notice that at the group level – or, in general, in small samples – stratifying the sample prior to randomization produces estimates with lower variance.

As one moves from column (1) to (4) the standard estimators worsen considerably. Note, however, that their RMSE's (in column (4) or (5)) though substantially rising do not exceed those of table 2 to a large extent. It is specifically the variation net of bias that has increased at the group level. Compared to the RMSE of the IV estimator the standard estimators still perform quite well. Albeit, their low RMSE's in column (5) should be taken with a grain of salt for different biases tend to cancel out due to special model constellations. This cannot be generalized. As above, in column (5) IV identifies

[16] At the group level, Z_2 dominates the selection equation because individual Z_1's aggregated to the group level are almost equal across all groups and, consequently, do not influence the selection process.

the mean effect on compliers instead of the effect on treated units, yet, its RMSE merely slightly increases with regard to columns (2) to (4).

Exploring the Potential of IV Estimators

Up to this point, results are generated under a certain simulation setup. Neither sample size nor the correlation between instrument and treatment indicator have been varied. For a thorough assessment of the relative performance of IV with respect to pure randomization it is necessary to perform a further simulation that varies these two parameters. The variables and parameters of the scenario reported in the fourth columns of tables 2 and 3 are selected and fixed. Table 4 presents ratios of the root mean squared error of IV and experimental estimates for certain correlations and number of groups.

As expected, IV produces more precise estimates as the correlation and the relative sample size increase. At the individual level, for a reasonable correlation of 0.3, the observational study should comprise ten times as many groups as the randomized experiment to generate a more efficient IV estimator. At the group level, the observational study should be at least 15 times as large as the group level experiment for the same correlation of instrument and treatment participation. Holding the relative sample size of the observational study at ten times as many groups a sufficient instrumental correlation would be around 0.4. This is already a high correlation but not completely utopian in practical applications.

Moreover, note that although the ratios of RMSE's at the group level are infinitely large for low correlations and low sample sizes, they diminish faster than they do at the individual level as correlations rise. Break-even points, i.e. points where the ratios are approximately 1 or less, are bold faced in the table; in general, they are later at the group level than at the individual level indicating that IV suffers more from the grouped structure than a randomized experiment. Only for the lowest number of groups in the first column reaches group-level IV its break-even point earlier. In all other columns the break-even point is reached for a slightly higher correlation.

Table 4: **IV Versus Experiment.**

	Number of groups in observational study				
Correlation	40	80	120	200	300
Individual Level					
0.105	11.740	5.094	4.003	2.794	2.196
0.199	5.419	2.734	1.964	1.502	1.181
0.301	3.454	1.847	1.354	**0.993**	**0.788**
0.401	2.705	1.355	**0.988**	0.746	0.586
0.502	2.042	1.104	0.800	0.621	0.471
0.601	1.728	**0.915**	0.695	0.511	0.406
0.703	1.469	0.762	0.564	0.418	0.348
0.799	1.214	0.669	0.536	0.380	0.298
0.904	**1.057**	0.579	0.451	0.322	0.267
0.991	0.961	0.521	0.416	0.305	0.226
Group Level					
0.100	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$
0.200	$+\infty$	$+\infty$	$+\infty$	$+\infty$	2.030
0.303	$+\infty$	$+\infty$	2.035	1.379	**0.987**
0.403	$+\infty$	1.797	1.233	**0.892**	0.708
0.500	$+\infty$	1.179	**0.914**	0.703	0.569
0.602	1.459	**0.961**	0.723	0.558	0.455
0.703	1.135	0.736	0.609	0.448	0.367
0.802	**0.958**	0.648	0.496	0.388	0.316
0.910	0.818	0.565	0.442	0.324	0.259
0.941	0.789	0.537	0.427	0.309	0.260

The table reports the ratio of root mean squared errors of the IV over the experimental estimate. The number of groups in the experimental setting is 20. Number of iterations: 2000.

5 Conclusion

This paper performs simulations in order to assess several standard estimation strategies such as the cross-sectional differences between treated and untreated units, before-after comparisons, and, specifically, instrumental variable estimators as a prime example of a quasi-experimental estimation strategy. These are compared to conventional randomized experiments under the assumption that experiments generally suffer from small sample problems. Therefore, they rely on markedly less observations than observational studies.

Standard estimators perform well as long as somewhat restrictive assumptions on the selection process are satisfied. In practical applications, it is typically difficult to justify the applicability of these assumptions. Therefore, randomized controlled experiments often provide the only credible counterfactual control group. However, situations are conceivable – particularly in social sciences – where randomized trials reach their limits. For instance, non-compliance, attrition, or randomization bias are well-known hazards of any experiment.[17] Focus here is rather on the problems caused by the small sample size typical for experiments which might set even more severe limits to evaluation. In this case, randomization might lose its persuasiveness for it cannot be expected to achieve balance of *all* relevant covariates between the treatment and control group.

Specifically, small sample sizes arise if randomization occurs at the group level and/or if cost considerations prevent analysts from establishing a large scale experiment. Therefore, alternatives should be considered as well. Instrumental variable estimation as a quasi-experimental technique might be a helpful device to circumvent the small sample problem and open the field for less costly large scale observational studies if a good instrument is available. The simulation results suggest that correlations of around 0.3 to 0.4 can be considered to characterize a good instrument if the observational study comprises ten times more observations than a corresponding randomized experiment. In practice, one might even encounter ratios larger than 10 which would thus allow to utilize instruments with lower correlations. Moreover, contaminations of randomized experiments – especially

[17] Since these problems are disregarded in our simulations they show randomized experiments in a favorable light. Other fundamental objections might be of ethical nature since treatments that produce positive effects are withheld the control group.

at the group level – would also be avoided in observational studies.

Albeit, IV estimation yields inconsistent estimates in case treatment effects are heterogeneous and individuals or groups decide whether to undergo treatment upon their true effects. In this case, IV identifies the mean effect of treatment on compliers, i.e. the local average treatment effect. Thus, it would answer the question of how large the treatment effect would be if the binary instrument Z were increased from 0 to 1, for example, if more treatment sites were established such that some individuals or groups had a shorter distance to their site. This measure would only affect compliers while always- and never-takers would be unaffected. From this point of view, LATE might give answers to policy relevant questions, too.[18] Nevertheless, it seems fairly unlikely that individuals know their own treatment effects in advance; in contrast, it seems more probable that they have to make their participation decision upon some sort of expected gains.

In sum, if a randomized experiment is infeasible because of practical reasons or because it would not provide enough observations, observational studies are not necessarily a contemptible alternative. They often contain valuable and detailed information that might still help to identify causal relationships. On the other hand, absent randomization bias and systematic attrition or noncompliance, randomized controlled experiments are the most convincing evaluation approach as long as a sufficient number of units are involved in the trial.

[18] See Angrist (1990), Angrist & Krueger (1991), and Imbens & Angrist (1994) for examples and a formal discussion.

References

Angrist, Joshua D. (1990) "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records", *American Economic Review*, 80: 313-36.

Angrist, Joshua D., Guido W. Imbens & Donald B. Rubin (1996) "Identification of Causal Effects Using Instrumental Variables", *Journal of the American Statistical Association*, 91: 444-72, with discussion by J.J. Heckman, R.A. Moffitt, J.M. Robins & S. Greenland, & R.P. Rosenbaum.

Angrist, Joshua D. & Alan B. Krueger (1991) "Does Compulsory Schooling Attendance Affect Schooling on Earnings?", *Quarterly Journal of Economics*, 106: 979-1014.

Angrist, Joshua D. & Alan B. Krueger (1999) "Empirical Strategies in Labor Economics", *Handbook of Labor Economics*, vol. 3, edited by Orely Ashenfelter and David Card. New York, NY: North-Holland, 1999 (forthcoming).

Bound, John, David A. Jaeger & Regina M. Baker (1995) "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variables is Weak", *Journal of the American Statistical Association*, 90: 443-50.

Bowden, Roger J. & Darrell A. Turkington (1984) *Instrumental Variables*, Cambridge: Cambridge University Press.

Burtless, Gary (1995) "The Case for Randomized Field Trials in Economic and Policy Research", *Journal of Economic Perspectives*, 9: 63-84.

Fisher, R.A. (1935) *The Design of Experiments*, Edinburgh: Oliver & Boyd.

Hall, Alastair R., Glenn D. Rudebusch & David W. Wilcox (1996) "Judging Instrument Relevance in Instrumental Variable Estimation", *International Economic Review* 37: 283-98.

Heckman, James J. (1996) "Randomization as an Instrumental Variable", *Review of Economics and Statistics*, 77(2): 336-41.

Heckman, James J. (1997) "Instrumental Variables, A Study of Implicit Behavioral Assumptions Used In Making Program Evaluations", *Journal of Human Resources*, 32: 441-62.

Heckman, James J., H. Ichimura & P. Todd (1997) "Matching as an Econometric Estimator: Evidence from Evaluating a Job Training Program", *Review of Economic Studies*, 64: 605-54.

Heckman, James J., Robert J. Lalonde & Jeffrey Smith (1999) "The Economics and Econometrics of Active Labor Market Programs", *Handbook of Labor Economics*, Volume 3, Chapter 31, edited by Orley Ashenfelter and David Card. New York, NY: North-Holland.

Heckman, James J. & Jeffrey A. Smith (1995) "Assessing the Case for Social Experiments", *Journal of Economic Perspectives*, 9: 85-110.

Heckman, J.J & R. Robb, Jr. (1985) "Alternative Methods for Evaluating the Impact of Interventions." In *Longitudinal Analysis of Labor Market Data*, ed. J. Heckman & B. Singer. New York: Cambridge University Press.

Imbens, Guido W. & Joshua D. Angrist (1994) "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 62: 446-475.

Kloek, T. (1981) "OLS Estimation in a Model Where a Microvariable Is Explained by Aggregates and Contemporaneous Disturbances Are Equicorrelated", *Econometrica*, 49: 205-07.

Manski, Charles F. (1990) "Nonparametric Bounds on Treatment Effects", *American Economic Review*, 80(2): 319-23.

Manski, Charles F. (1995) *Identification Problem in the Social Sciences*, Cambridge, MA: Harvard University Press.

Moulton, Brent R. (1986) "Random Group Effects and the Precision of Regression Estimates", *Journal of Econometrics*, 32: 385-97.

Moulton, Brent R. (1990) "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units", *Review of Economic and Statistics*, 71: 334-38.

Murray, David M. (1998) *Design and Analysis of Group-Randomized Trials*, Oxford: Oxford University Press.

Rosenbaum, Paul R. & Donald B. Rubin (1983) "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70: 41-55.

Rosenbaum, Paul R. & Donald B. Rubin (1984) "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score", *Journal of the American Statistical Association*, 79: 516-24.

Rosenbaum, Paul R. & Donald B. Rubin (1985) "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score", *The American Statistician*, 39: 33-38.

Rubin, Donald B. (1973) "Matching to Remove Bias in Observational Studies", *Biometrics*, 29: 159-83, Printer's correction (1974), 30: 728.

Rubin, Donald B. (1974) "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, 66: 688-701.

Rubin, Donald B. (1986) "What Ifs Have Causal Answers?", *Journal of the American Statistical Association*, 81: 961-62.

Schmidt, Christoph M., Rob Baltusen & Rainer Sauerborn (1999) "Evaluation of Community-Based Interventions: Group-Randomization, Limits and Alternatives", *Discussion paper series* no. 281, Department of Economics, University of Heidelberg.

Shore-Sheppard, Lara (1996) "The Precision of Instrumental Variables Estimates With Grouped Data", *Industrial Relations Section Working Paper 374*, Princeton University.

IZA Discussion Papers

No	Author(s)	Titel	Area	Date
181	E. Wasmer Y. Zenou	Space, Search and Efficiency	2	8/00
182	M. Fertig C. M. Schmidt	Discretionary Measures of Active Labor Market Policy: The German Employment Promotion Reform in Perspective	6	8/00
183	M. Fertig C. M. Schmidt	Aggregate-Level Migration Studies as a Tool for Forecasting Future Migration Streams	1	8/00
184	M. Corak B. Gustafsson T. Österberg	Intergenerational Influences on the Receipt of Unemployment Insurance in Canada and Sweden	3	8/00
185	H. Bonin K. F. Zimmermann	The Post-Unification German Labor Market	4	8/00
186	C. Dustmann	Temporary Migration and Economic Assimilation	1	8/00
187	T. K. Bauer M. Lofstrom K. F. Zimmermann	Immigration Policy, Assimilation of Immigrants and Natives' Sentiments towards Immigrants: Evidence from 12 OECD-Countries	1	8/00
188	A. Kapteyn A. S. Kalwij A. Zaidi	The Myth of Worksharing	5	8/00
189	W. Arulampalam	Is Unemployment Really Scarring? Effects of Unemployment Experiences on Wages	3	8/00
190	C. Dustmann I. Preston	Racial and Economic Factors in Attitudes to Immigration	1	8/00
191	G. C. Giannelli C. Monfardini	Joint Decisions on Household Membership and Human Capital Accumulation of Youths: The role of expected earnings and local markets	5	8/00
192	G. Brunello	Absolute Risk Aversion and the Returns to Education	5	8/00
193	A. Kunze	The Determination of Wages and the Gender Wage Gap: A Survey	5	8/00
194	A. Newell F. Pastore	Regional Unemployment and Industrial Restructuring in Poland	4	8/00
195	F. Büchel A. Mertens	Overeducation, Undereducation, and the Theory of Career Mobility	5	9/00

196	J. S. Earle K. Z. Sabirianova	Equilibrium Wage Arrears: A Theoretical and Empirical Analysis of Institutional Lock-In	4	9/00
197	G. A. Pfann	Options to Quit	1	9/00
198	M. Kreyenfeld C. K. Spiess G. G. Wagner	A Forgotten Issue: Distributional Effects of Day Care Subsidies in Germany	3	9/00
199	H. Entorf	Rational Migration Policy Should Tolerate Non-Zero Illegal Migration Flows: Lessons from Modelling the Market for Illegal Migration	1	9/00
200	T. Bauer G. S. Epstein I. N. Gang	What are Migration Networks?	1	9/00
201	T. J. Dohmen G. A. Pfann	Worker Separations in a Nonstationary Corporate Environment	1	9/00
202	P. Francois J. C. van Ours	Gender Wage Differentials in a Competitive Labor Market: The Household Interaction Effect	5	9/00
203	J. M. Abowd F. Kramarz D. N. Margolis T. Philippon	The Tail of Two Countries: Minimum Wages and Employment in France and the United States	5	9/00
204	G. S. Epstein	Labor Market Interactions Between Legal and Illegal Immigrants	1	10/00
205	A. L. Booth M. Francesconi J. Frank	Temporary Jobs: Stepping Stones or Dead Ends?	1	10/00
206	C. M. Schmidt R. Baltussen R. Sauerborn	The Evaluation of Community-Based Interventions: Group-Randomization, Limits and Alternatives	6	10/00
207	C. M. Schmidt	Arbeitsmarktpolitische Maßnahmen und ihre Evaluierung: eine Bestandsaufnahme	6	10/00
208	J. Hartog R. Winkelmann	Dutch Migrants in New Zealand: Did they Fare Well?	1	10/00
209	M. Barbie M. Hagedorn A. Kaul	Dynamic Effciency and Pareto Optimality in a Stochastic OLG Model with Production and Social Security	3	10/00
210	T. J. Dohmen	Housing, Mobility and Unemployment	1	11/00
211	A. van Soest M. Das X. Gong	A Structural Labour Supply Model with Nonparametric Preferences	5	11/00

212	X. Gong A. van Soest P. Zhang	Sexual Bias and Household Consumption: A Semiparametric Analysis of Engel Curves in Rural China	5	11/00
213	X. Gong A. van Soest E. Villagomez	Mobility in the Urban Labor Market: A Panel Data Analysis for Mexico	1	11/00
214	X. Gong A. van Soest	Family Structure and Female Labour Supply in Mexico City	5	11/00
215	J. Ermisch M. Francesconi	The Effect of Parents' Employment on Children's Educational Attainment	5	11/00
216	F. Büchel	The Effects of Overeducation on Productivity in Germany — The Firms' Viewpoint	5	11/00
217	J. Hansen R. Wahlberg	Occupational Gender Composition and Wages in Sweden	5	11/00
218	C. Dustmann A. van Soest	Parametric and Semiparametric Estimation in Models with Misclassified Categorical Dependent Variables	1	11/00
219	F. Kramarz T. Philippon	The Impact of Differential Payroll Tax Subsidies on Minimum Wage Employment	5	11/00
220	W. A. Cornelius E. A. Marcelli	The Changing Profile of Mexican Migrants to the United States: New Evidence from California and Mexico	1	12/00
221	C. Grund	Wages as Risk Compensation in Germany	5	12/00
222	W.P.M. Vijverberg	Betit: A Family That Nests Probit and Logit	7	12/00
223	M. Rosholm M. Svarer	Wages, Training, and Job Turnover in a Search-Matching Model	1	12/00
224	J. Schwarze	Using Panel Data on Income Satisfaction to Estimate the Equivalence Scale Elasticity	3	12/00
225	L. Modesto J. P. Thomas	An Analysis of Labour Adjustment Costs in Unionized Economies	1	12/00
226	P. A. Puhani	On the Identification of Relative Wage Rigidity Dynamics: A Proposal for a Methodology on Cross-Section Data and Empirical Evidence for Poland in Transition	4/5	12/00

227	L. Locher	Immigration from the Eastern Block and the former Soviet Union to Israel: Who is coming when?	1	12/00
228	G. Brunello S. Comi C. Lucifora	The College Wage Gap in 10 European Countries: Evidence from Two Cohorts	5	12/00
229	R. Coimbra T. Lloyd-Braga L. Modesto	Unions, Increasing Returns and Endogenous Fluctuations	1	12/00
230	L. Modesto	Should I Stay or Should I Go? Educational Choices and Earnings: An Empirical Study for Portugal	5	12/00
231	G. Saint-Paul	The Economics of Human Cloning	5	12/00
232	E. Bardasi M. Francesconi	The Effect of Non-Standard Employment on Mental Health in Britain	5	12/00
233	C. Dustmann C. M. Schmidt	The Wage Performance of Immigrant Women: Full-Time Jobs, Part-Time Jobs, and the Role of Selection	1	12/00
234	R. Rotte M. Steininger	Sozioökonomische Determinanten extremistischer Wahlerfolge in Deutschland: Das Beispiel der Europawahlen 1994 und 1999	3	12/00
235	W. Schnedler	Who gets the Reward? An Empirical Exploration of Bonus Pay and Task Characteristics	5	12/00
236	R. Hujer M. Caliendo	Evaluation of Active Labour Market Policy: Methodological Concepts and Empirical Estimates	6	12/00
237	S. Klasen I. Woolard	Surviving Unemployment without State Support: Unemployment and Household Formation in South Africa	3	12/00
238	R. Euwals A. Börsch-Supan A. Eymann	The Saving Behaviour of Two Person Households: Evidence from Dutch Panel Data	5	12/00
239	F. Andersson K. A. Konrad	Human Capital Investment and Globalization in Extortionary States	5	01/01
240	W. Koeniger	Labor and Financial Market Interactions: The Case of Labor Income Risk and Car Insurance in the UK 1969-95	5	01/01

241	W. Koeniger	Trade, Labor Market Rigidities, and Government-Financed Technological Change	2	01/01
242	G. Faggio J. Konings	Job Creation, Job Destruction and Employment Growth in Transition Countries in the 90's	4	01/01
243	E. Brainerd	Economic Reform and Mortality in the Former Soviet Union: A Study of the Suicide Epidemic in the 1990s	4	01/01
244	S. M. Fuess, Jr. M. Millea	Pay and Productivity in a Corporatist Economy: Evidence from Austria	5	01/01
245	F. Andersson K. A. Konrad	Globalization and Human Capital Formation	5	01/01
246	E. Plug W. Vijverberg	Schooling, Family Background, and Adoption: Does Family Income Matter?	5	01/01
247	E. Plug W. Vijverberg	Schooling, Family Background, and Adoption: Is it Nature or is it Nurture?	5	01/01
248	P. M. Picard E. Toulemonde	The Impact of Labor Markets on Emergence and Persistence of Regional Asymmetries	2	01/01
249	B. M. S. van Praag P. Cardoso	"Should I Pay for You or for Myself?" The Optimal Level and Composition of Retirement Benefit Systems	3	01/01
250	T. J. Hatton J. G. Williamson	Demographic and Economic Pressure on Emigration out of Africa	1	01/01
251	R. Yemtsov	Labor Markets, Inequality and Poverty in Georgia	4	01/01
252	R. Yemtsov	Inequality and Income Distribution in Georgia	4	01/01
253	R. Yemtsov	Living Standards and Economic Vulnerability in Turkey between 1987 and 1994	4	01/01
254	H. Gersbach A. Schniewind	Learning of General Equilibrium Effects and the Unemployment Trap	3	02/01
255	H. Gersbach A. Schniewind	Product Market Reforms and Unemployment in Europe	3	02/01
256	T. Boeri H. Brücker	Eastern Enlargement and EU-Labour Markets: Perceptions, Challenges and Opportunities	2	02/01

257	T. Boeri	Transition with Labour Supply	4	02/01
258	M. Rosholm K. Scott L. Husted	The Times They Are A-Changin': Organizational Change and Immigrant Employment Opportunities in Scandinavia	1	02/01
259	A. Ferrer-i-Carbonell B. M.S. van Praag	Poverty in the Russian Federation	4	02/01
260	P. Cahuc F. Postel-Vinay	Temporary Jobs, Employment Protection and Labor Market Performance	1/3	02/01
261	M. Lindahl	Home versus School Learning: A New Approach to Estimating the Effect of Class Size on Achievement	5	02/01
262	M. Lindahl	Summer Learning and the Effect of Schooling: Evidence from Sweden	5	02/01
263	N. Datta Gupta N. Smith	Children and Career Interruptions: The Family Gap in Denmark	5	02/01
264	C. Dustmann	Return Migration, Wage Differentials, and the Optimal Migration Duration	1	02/01
265	M. Rosholm M. Svarer	Structurally Dependent Competing Risks	1	02/01
266	C. Dustmann O. Kirchkamp	The Optimal Migration Duration and Activity Choice after Re-migration	1	02/01
267	A. Newell	The Distribution of Wages in Transition Countries	4	03/01
268	A. Newell B. Reilly	The Gender Pay Gap in the Transition from Communism: Some Empirical Evidence	4	03/01
269	H. Buddelmeyer	Re-employment Dynamics of Disabled Workers	3	03/01
270	B. Augurzky C. M. Schmidt	The Evaluation of Community-Based Interventions: A Monte Carlo Study	6	03/01