Journal of Experimental Criminology (2006) 2:375–391
DOI: 10.1007/s11292-006-9016-7

Improving impact evaluations through randomised experiments: The challenge of the National Research Council report for European criminology

MARTIN KILLIAS
Ecole des sciences criminelles, University of Lausanne, CH-1015 Lausanne, Switzerland
E-mail: martin.killias@unil.ch

Abstract. The National Research Council (NRC) report on Improving Evaluation of Anticrime Programs presents and discusses a wide array of techniques of evaluation. Although recognising the very high internal validity of randomised experiments, it considers, under certain conditions, quasi-experiments and observational studies as equally valid approaches. This conclusion is critically reviewed from a European perspective, where only a few randomised trials have been realised so far. It is argued that many critiques routinely addressed to randomised experiments, such as ethical concerns or low acceptance among practitioners, are either unfounded or can be adequately dealt with through imaginative adjustments. On the other hand, randomised controlled trials need to take the challenge of broadening the perspective, especially by looking at long-term effects that no other method can consider with comparable internal validity. Other recommendations include using innovative measures of re-offending, considering dynamic rather than static criteria of re-offending, and looking, beyond re-offending, at rehabilitation in other areas of life. Particular challenges are the possible placebo effects that evaluators in criminal justice have not yet found appropriate ways to deal with.

Key words: ethics, long-term effects, placebo effects, randomization, recidivism, rehabilitation

Background

The National Research Council's (NRC's) report on "Improving Evaluation of Anticrime Programs" (2005) comes as a real challenge, perhaps even more so for Europeans than for Americans. Long dominated by legal scholars interested in issues of crime policy, criminological evaluations in Europe rarely met international standards. In recent years, however, criminology and evaluations have changed not only in England, but on the continent as well. Recent developments have been less acknowledged on the west side of the Ocean, however, a fact well documented by the quasi-absence of European criminology in American bibliographies, beyond routinely paid tribute to historic figures such as Lombroso, Marx, Durkheim and Weber. Over the past generation, however, many criminologists have entered the field who happened to be trained, either as a principal or as a minor subject, in social sciences. This has also been reflected also in evaluations of anti-crime and correctional programmes. However, many high-quality evaluations conducted in Europe continue to be ignored in American reviews.[1] In this paper we shall discuss randomised trials in the field of criminal justice throughout Europe, pointing to shortcomings and strengths in light of the recommendations presented

in the NRC report. With few exceptions, we shall focus on impact evaluations of sanctions and treatment programmes aimed at reducing offending.

Randomised experiments in the NRC report

The NRC report pays, on several occasions, tribute to randomised experiments that "*should be favored in situations where it is likely that they can be implemented with integrity and will yield useful results*" (p. 4). However, the final conclusion is that the choice of the method of evaluation should be the result of "*...careful deliberation about which evaluation design is likely to yield the most useful and relevant information...rather than generalizations about the relative superiority of one method over another*" (p. 40). Further doubts appear in the statement that "*in some cases, randomization is not acceptable for political or ethical reasons*" (p. 39), as for example in "experiments" on the effectiveness of the death penalty—an obviously true but irrelevant reservation, as will be shown later in this paper. Other examples of reservations are the statement that attrition can be a major threat to randomised experiments (p. 36), or that observational studies may have higher external validity (p. 36) at lower costs (p. 36). All three statements are presented without further illustrations on how they have been reached. As we shall argue, no one is necessarily and always true.

The "natural" European country to conduct randomised trials would probably be England, given that British criminology is firmly embedded in social sciences and that hundreds of scholars are dealing with evaluations of all sorts of programmes there. Indeed, after the landmark studies by Mannheim and Wilkins (1955) and Wilkins' writings on this subject (1969), experimental studies became increasingly popular during the early 1970s. Unfortunately, a major drawback came when the evaluation of a major institution for juveniles (Kingswood Training School) raised serious problems, due to the way randomisation affected the institution and decisions to place juvenile offenders in that setting (Cornish 1987). It seems that the lesson learned from that experiment had not been that controlled trials need, eventually, to be improved and made compatible with practical requirements, but that they are rarely feasible and should be replaced by an alternative technique of evaluation (Tilly 2000). Over the following decades, it seems that British criminology widely abandoned controlled trials, with few exceptions such as the Kent Intensive Supervision and Support Programme (ISSP) trial (Little et al. 2004). Recently, however, there seems to have been a turn-around, governmental policy makers having discovered "cost efficiency" and, thus, the need for conclusive evaluations. How far that will lead remains to be seen.

On the continent, things were hardly more encouraging. Given the legal background of virtually all policy makers, it was all too easy to dismiss randomisation for being "legally" or "ethically" unacceptable, without further consideration of such issues. In addition, policy makers and managers involved in innovative correctional programmes were usually so convinced about their beneficial effects that no need was felt to evaluate them in a way that might

have been able to show negative outcomes. As a rule, evaluations were not set up to promote learning from experience but to legitimise after the fact programmes adopted by policy makers, or to provide "scientific" support for initiatives to expand them to the entire criminal justice system.

The debate on randomisation

Learning from America—and beyond

The increasing popularity of randomised controlled trials in America and in some other parts of the world, as evidenced by the growing attention and prestige of networks such as the Campbell Collaboration Crime and Justice Group, have produced some ironic consequences in Europe. Usually, scientific meetings include increasing numbers of sessions devoted to meta-analyses and experimental research, often introduced by well-known "stars" in this field from England and America. If the interest documented in such initiatives certainly deserves to be welcomed, the way this new research tradition is being presented often induces people in Europe to consider this as an "American" approach that may be admirable but that will not have any practical bearing on their own work. At the same time, experimental research in Europe, usually unknown to American scholars, will be the more overlooked by Europeans, who always had far greater difficulties to learn from each other than from the other side of the Ocean.[2] This is particularly unfortunate, because randomised experiments conducted in one neighbouring country might be a most valuable source of inspiration on how such a research design could be made feasible. Given the many reservations and obstacles to overcome, this practical know-how is probably even more critical than familiarity with sophisticated methods of data analysis. Looking at this state of affairs it is, despite its obvious merits in discussing all available methods of evaluations in criminal justice, not impossible that the NRC report may come as a drawback in the European context. Discovering that a body of most distinguished American (and, thus, leading) evaluation experts recommend considering softer designs as equally valid alternatives to randomised controlled trials, may induce European sceptics to resist, even more persuasively, pressure to increase randomised trials. In this context the critical overtones in the American debate, such as the critique of alternative methods by Berk (2005), are likely to be overheard in Europe.

Improving external validity

Seen from America, these continental parochialisms may look trivial, but they go along with an unfortunate lack of replication of results found within the United States of America. Indeed, the high internal validity of randomised controlled trials does not imply an equally high external validity, as rightly stated in the NRC report

(p. 41). Results obtained in the USA cannot automatically be generalised to the rest of the world, particularly when American experts are reluctant about generalising outcomes across their own country. Of course, the external validity of European studies is no less questionable. If, however, certain interventions are followed by similar results on both sides of the Ocean, our confidence may increase that the intervention at stake is likely to produce similar outcomes when replicated in different contexts.

Obviously, there will be no disagreement about this. However, the question is whether observational studies really offer better external validity at lower costs, as stated in the NRC report (p. 40). Of course, observational (and also quasi-experimental) studies are relatively easy to replicate in various sites, but the question is, at what price? If judges systematically sentence to prison those defendants with the worst prospects of re-offending, whereas the best-risk offenders tend, everywhere, to be sent on an "alternative" programme, the likely outcome will obviously be that prisoners everywhere will have the highest rates of re-offending. Under such circumstances, a meta-analysis is likely to reproduce (and multiply) findings of highly questionable internal validity rather than increasing external validity. It should, perhaps, be better recognised that external validity cannot be increased by generalising from studies of poor internal validity.

Policy makers' interest in randomised controlled trials

In the future it will be very important to develop evaluation standards in the field of research on re-offending in order to improve the quality of trials. Moreover, randomised controlled trials ought to be preferred not only by researchers but also by policy makers. Only experimental research designs can establish the relative effects of different interventions on re-offending or any other outcomes. Not knowing the answer to this question means that we cannot tell whether taxpayers' money has been spent efficiently or not. In comparison with quasi-experimental research designs, randomised trials are, according to our experience, not necessarily more expensive, since far fewer independent (control) variables need to be collected over time in order to assess programme effects.[3] In other words, randomised controlled trials offer higher quality at often far lower costs. Interestingly, this aspect has hardly been touched on in the NRC report (except in a short remark on p. 40).

Policy makers may legitimately award less importance to randomised trials whenever interventions are followed by extreme rates of success or failure. As in the medical field, interventions sometimes produce overwhelming effects or produce very strong undesirable side effects that oblige the programmes to be stopped immediately.[4] Under such conditions, it would be absurd to insist on a randomised trial. Normally, however, the effects of interventions in criminal justice are somewhere between these extremes, so that the question of relative success or failure, or potential of improvement through alternative approaches, cannot be ruled out.

Ethical reservations

Sceptics tend to reply to the call for more controlled trials by pointing to ethical, practical or legal difficulties with such research designs. Apparently, such reservations have also inspired the NRC report (p. 39), although the two examples offered may be questionable. Evaluating the death penalty through a randomised experiment is absurd, not so much because, for those executed, there will be no follow-up period, but, obviously, because there are interventions that are unethical whatever the evaluation method adopted. It is true also that, in the case of serious offenders, randomising between some form of confinement and non-custodial sanctions would be unethical, but, again, because of ethical problems in letting dangerous individuals go free rather than because of randomisation as such. There have been randomised trials involving very serious offenders in Germany, where subjects were randomly assigned either to treatment in a so-called social therapy unit (Sozialtherapeutische Anstalt) or in the ordinary prison regime (Ortmann 2000). Therefore, ethical reservations are directed, in these instances, at problematic interventions rather than at the method of evaluation, particularly if interventions go along with irreversible damage that is not balanced by any known benefits. However, even in cases of most severe interventions, as, for example, when offenders with a long record of serious sexual violence are subject to chemical (i.e., reversible) or (irreversible) surgical castration, ethical reservations should not bar such programmes from being tested. According to a recent meta-analysis of sex offender therapies, Lösel and Schmucker (2005) concluded that such interventions are, by far, more efficient at preventing re-offending than any other approaches, including cognitive behaviour treatment (that was found promising but far less efficient). At least in the case of offenders who agree to such interventions, their benefits and risks should be balanced against the suffering by victims that will eventually be prevented, as well as against the pain inflicted to offenders who otherwise would be likely to spend many years behind bars.

Finally, ethical arguments seem to be quite odd so long as no evidence has shown that "new" sanctions or programmes produce better results than traditional ones, or that they are at least not damaging. No one encourages pharmaceutical firms to sell new products before adequate testing through randomised controlled trials. Why should a new correctional programme be "sold" to participants when its effects have not been adequately tested, simply because a few correctional specialists argue, more or less convincingly, that it may most likely be beneficial?

Concerns about low acceptance of randomisation

Since we launched two experimental trials in the field of correction, in 1993 and in 2000 in Switzerland, our experience with correctional services, convicted offenders participating in these programmes and policy makers has been that random assignment has many advantages, also, for staff and decision makers operating in the field. The Swiss prison vs community service experiment, for example, has

shown that randomisation was quite well accepted in both groups, despite some evidence to the contrary in the literature (Erez 1985). Therefore, random assignment may often be easier to justify than any kind of choice on the grounds of personal characteristics, merits or institutional constraints, particularly if the number of candidates exceeds the planned capacity of the experimental group (Weisburd 2000).

Serious practical problems can arise whenever practitioners are highly committed to a programme, as they should be (Petrosino and Soydan 2005), and if random assignment is rigidly applied without due consideration of practical concerns. There are often programmes that are designed only for the treatment of individuals having certain characteristics. In such cases it is good practice to screen subjects to assess their eligibility, as for example in the case of the Sozialtherapeutische Anstalten in Germany (Ortmann 1994, 2000).[5] Even if practitioners thus have the possibility of eliminating from the treatment group those subjects that are not suitable, they may have strong reservations whenever a "particularly needy" subject is being assigned to the control group (Little et al. 2004). In order to defend the trial's integrity, an excellent method, already suggested by Wilkins (1969), is to admit a certain (pre-fixed) number of subjects before any randomisation takes place. In the case of the Swiss community service vs prison experiment, social workers were allowed to admit up to 25% of subjects before (i.e., without) randomisation. These subjects were, of course, not comparable to the two randomised groups. For this reason they were kept identifiable and were analysed separately. In the end, this procedure has not only reduced the temptation to "cheat" during randomisation (and, thus, increased the experiment's integrity) but has also contributed to increase the acceptability of randomisation among practitioners and to prevent covert opposition to the trial, as in the case of Kingswood Training School, described by Cornish (1987); further, this approach also offered the rather unique chance to study the profile and outcome of subjects the practitioners had defined before as "particularly needy".[6] Improving acceptance of a programme or of random assignment of subjects can also help to control attrition, a notorious problem of many randomised experiments (NRC report, p. 36). Eisner and Ribeaud (2006) used several imaginative means to reduce panel attrition, particularly among minority parents, in the Zurich Intervention and Prevention Project, an experimental longitudinal study of 1,240 children.

Legislative action to legalise trials

In some countries it may be wise to remove legal obstacles by appropriate legislative actions. For example, the Swiss parliament adopted, in 1971, an amendment to the penal code (section 397 bis par. 4) allowing the Government to introduce, on an experimental basis, i.e., for a limited number of offenders and for a limited period of time, innovative sanctions and correctional arrangements beyond those provided for by the penal code. Under this law, offenders who are offered the chance of serving their term in an "innovative" programme may, at any time, refuse and claim to be

treated "according to the law" (i.e., serving their term in prison, as a rule); however, no one is entitled to claim to become part of an experiment that is, by essence, limited in scope. Therefore, no legal obstacle precludes randomisation among those who volunteer and are eligible for an "experimental" sanction or programme. Given that new sanctions had been introduced in many other countries as a temporary and more or less "experimental" arrangement, similar provisions for high-quality evaluations should have been no less feasible. It is, therefore, hard to imagine that randomisation should not be legally feasible whenever electronic monitoring, or whatever "alternative", is being offered to convicted defendants as an "experimental" arrangement beyond the penal code. Of course, such programmes are necessarily always limited to "volunteers", but, given the high demand for such arrangements among convicted defendants, doubts about external validity (regarding non-volunteers) may not be all too pressing.

Alternatives to randomised experiments

If the unit of analysis is a city, a region or an entire country, randomised controlled trials are not feasible, except in cases where the number of units is such that different groups of reasonable size can be selected, as in the Minneapolis Hot Spots Experiment (Sherman and Weisburd 1995), in the Nuremberg kindergarten project (Lösel et al. 2006) or in the Zurich Intervention and Prevention Project (Eisner and Ribeaud 2005). Under "normal" conditions, where some policy or intervention change is planned at a few sites, however, a different approach seems more appropriate, such as non-randomised comparisons of sites with and without change over time, as recommended in the NRC report (p. 37). For example, a certain change can be implemented in a group of sites G-1 at t-1 and evaluated at t-2, t-3, t-4, etc., whereas other sites (grouped to G-2) start with the new approach at t-2, followed by a third group (G-3) at t-3, etc. If the outcome (measured at t-2, t-3, t-4 etc.) is consistently the same in all groups of sites *after* implementation of the change, it may be hard to attribute the apparent effect to any third variable that remained uncontrolled. According to this method, and thanks to the fact that community service has been introduce gradually in 22 out of Switzerland's 26 cantons over a period extending from 1991 to 1999, it has been possible to show a net-widening effect of this new sanction in cases of theft.[7] A rather rare form of evaluation is to introduce a certain intervention (a new law or any sort of measure) at t-1 and then see whether the dependant variable (i.e., in general, offending) varies at t-2. Unlike in simple before-and-after comparisons (as discussed in the NRC report on p. 37), one can make such a test more convincing by revoking the change (the law or measure) at t-3 and then see whether some contrary effect can be observed; if the data collected at t-4 show a change in the opposite (presumably undesirable) direction, one can, at t-5, re-introduce the same law or measure again and see whether the same effect as the one observed at t-2 will appear again at t-6. The effect of a law making the use of safety belts in cars compulsory has been evaluated in Switzerland through an (unwanted) experiment of this sort (Killias

1985). It showed that behaviour of drivers adjusted each time to changing legal regulations, showing a considerable "deterrent" effect of legal change (rather than punishment as such). Unfortunately, legislators can hardly legitimise this kind of trial-and-error approach under normal condition, which is the reason why such experiments are extremely rare in practice.[8]

A brilliant natural experiment was conducted during the 1960s in the Netherlands. Following an old tradition, the wedding of Princess (now Queen) Beatrix and Prince Claus von Amsberg had led Queen Juliana to enact a decree of pardon for defendants sentenced to a custodial sentence of no more than 14 days for offences committed prior to January 1st 1966. Those who had their immediate custodial sentences commuted to a suspended one did, obviously and with the exception of the date of offending, not differ from their less happy counterparts who had offended after that date and who had to serve their sentence in prison. Under these circumstances, it can reasonably be assumed that both groups were similar in terms of risks of re-offending. The evaluation by van der Werff (1979, 1981) showed that re-conviction rates did not differ between those sent to prison and those who had had their sentences commuted.[9] Given the unusually large sample and the long follow-up period (6 years), this experiment may still be considered one of the most significant tests of the effects of custodial and non-custodial sanctions on re-offending.

Improving impact evaluations beyond the NRC report

The review of controlled trials in this essay, as well as a systematic review of studies having compared re-offending rates following custodial and non-custodial ("alternative") sanctions (Villettaz et al. 2006), has allowed us to identify a number of shortcomings that might be relatively easy to overcome in the future, no matter in what country the evaluation is to take place. The NRC's report will be paid due attention to in the following discussion.

More randomised controlled trials are needed

If we look at evaluations in Europe in the field of criminal justice, the first and most obvious conclusion concerns the lack of controlled experiments that, for the time being, remain rare exceptions, often promoted by dedicated researchers or policy makers who, for whatever reason, are committed to "objective" results rather than to "proof" that their policies worked. The result of this situation is the impossibility to draw firm conclusions about the effects of custodial vs non-custodial sanctions, despite hundreds of evaluations conducted worldwide over this question (Smith et al. 2002; Villettaz et al. 2006). Therefore, randomisation should become a far more acceptable, if not standard, option for policy makers who mandate evaluations of any new forms of treatment or sanction. The obstacles that are routinely invoked are far less absolute than often claimed. Once the number of randomised experiments has increased, researchers and policy makers will probably learn how to overcome legal

and ethical obstacles in acceptable ways—everything is, in the end, a question of *how* rather than of *whether* controlled trials can be conducted.

Discovering long-term effects

The lack of randomised trials is only part of the problem. No less deplorable is the fact that, whatever the method adopted, follow-up periods in evaluations rarely extend beyond 2 years, presumably because policy makers need rapid "feedback". It should be recognised that randomised experiments make longer observation periods far more feasible. If subjects were, originally, randomly assigned to different conditions, their development over their entire lifespan can be studied without undue investment in time and resources. Quasi-experiments will, at best, allow one to control some variables that might have an impact on the choice of the type of intervention as well as on foreseeable outcomes, but will certainly not allow control of such influences on unpredicted long-term outcomes, such as unanticipated health problems. If, for example, subjects in the treatment group suffer later in their lives more often than those in the control group of cardio-vascular problems, as observed in the case of the Cambridge Somerville experiment (McCord 1990), it would, without randomisation, not be possible to rule out that, from the onset, candidates with more vulnerable health had been assigned disproportionately to the treatment group. Indeed, no checks of their heart condition had been performed at that time. Probably due to the unpopularity of randomisation, studies conducted, so far, in Europe never have extended to significant parts of subjects' later biographies. This is most unfortunate, since several evaluations conducted in America (McCord 1990; Schweinhart et al. 1993; Olds and Kitzman 1993) have shown that positive as well as negative effects are often far stronger in the long run than in the short run. It is most fortunate that few experimental studies involving school children have been started over the past years that may, one day, allow us to study long-term effects, such as the Zurich Intervention and Prevention Project by Eisner and Ribeaud (2005), or a study based on randomly selected pre-school facilities with 675 children aged about 4 to 5 years in the Nuremberg region (Germany) by Lösel et al. (2006). In European countries where population mobility (particularly across national and language barriers) remains relatively modest, long-term studies could be particularly fascinating, given the availability of many data in official records over extended periods of time. This is the case in many European countries, although the best-known example is Denmark, where huge databases covering entire biographies can be matched. Brennan and Mednick (1994) used such records to study re-convictions in an entire birth cohort (born in Copenhagen between 1944 and 1947).

Beyond official records as measures of "success" or "failure"

Despite alternative (and presumably more valid) measures of re-offending (such as self-reports), most studies do not include measures of re-offending beyond re-arrest

or re-conviction. Given the strong correlation between offending and victimisation, one might also validly consider, in evaluation research, a combination of self-report and victimisation questionnaires in order to assess effectiveness of programmes. If a programme is, indeed, successful at reducing offending rates, one should also be able to identify such an effect through reduced rates of victimisation. This is not trivial, since questions on victimisation provoke less resistance from subjects than self-report instruments. Evaluations of drug treatment programmes, such as heroin and methadone prescription in Switzerland and in German cities (Bonn, König 2002; Hamburg, Legge and Bathsteen 2000), have combined interviews covering self-reported delinquency and victimisation with police and conviction records. A methodological study on some 500 subjects (Aebi 2006) has documented reasonably high validity of all three methods to identify programme effects on the *prevalence* of offending. However, it turned out that variations in individual *incidence* rates (i.e., the frequency of offending) tend to be underestimated in official records. Based on these results, and in order to reduce costs, later evaluations (after 4 years) have used official records only and dropped interviews (Killias et al. 2005; Ribeaud 2004).[10]

Looking at relative improvement

In most studies re-offending has been measured through the prevalence of post-intervention re-convictions or re-arrests. Left alone that questionnaires of self-reported delinquency and/or victimisation were rarely used, the simple prevalence ("yes/no") of arrests or convictions after an intervention may mask important variations in the frequency of offending ("incidence rates") and relative improvement following different sanctions (Little et al. 2004). Depending on the population studied, convictions are not necessarily frequent and may, eventually, not allow one to observe sufficient variance in order to discover any sanction (or intervention) effect, especially if the sample is not large.[11] This is particularly true under the continental sentencing system where one global sentence is imposed for all (new) offences of which the defendant has been found guilty, rather than one sentence for each verdict, as under the Anglo-American system. If re-arrest data are used, this problem is less serious because police contacts are more frequent than re-convictions, one court appearance being eventually related to several new offences known by the police. However, survey data systematically allow one to observe far higher rates of re-offending than any official measures (Aebi 2006).

Some studies have shown that most offenders reduce offending rates after whatever type of intervention (Empey and Lubeck 1971; Killias et al. 2000a). Thus, the relevant question may be to what extent they improve differently by type of sanction. Therefore, it would be urgent to look in future studies at rates of improvement (or reductions in offending) rather than merely at "recidivism" as such. This is particularly true if samples are not very large, if "failure" rates are not very high (or not very different across groups) and if, as not unusual in such situations, subjects' pre-intervention offending rates were, despite randomisation,

higher in one group than in another.[12] If we compare pre- and post-intervention incidence rates (e.g., number of offences known to the police during the 2 years before and after the intervention), statistical power can be increased under such circumstances. It is somewhat surprising that this issue is, apparently, not raised in the NRC report. As rightly stated in the report (p. 43), increasing the sample size often raises practical difficulties or even produces adverse side effects, such as reducing the "dosage" of treatment (Weisburd et al. 1993).

Looking at rehabilitation beyond re-offending

In studies comparing custodial and non-custodial sanctions, lower re-offending rates among those sentenced to an "alternative" sanction were, whenever observed, usually attributed to the fact that these offenders were not separated from their work and family life and, therefore, may have had better opportunities to integrate after having served their sentence. However, the evidence is extremely limited in this respect (Lamb and Goertzel 1974; Killias et al. 2000a) and does not necessarily confirm this assumption, since almost all studies focus on re-offending (Israel and Chui 2006). Given the often extremely short duration of custodial sentences compared to "alternative" sanctions under European law, it seems unlikely that any lasting "prisonisation" effect may have been produced. In the case of randomised controlled trials, it would be easy to conduct later follow-up studies including, beyond measures of re-offending, any kind of indicators of social integration, as they can routinely be found, for example, in the files of income revenue services. The files of such services routinely collect data on family disruption, unemployment, welfare payments, debts, revenues and resources. Such data would be highly relevant in assessing any negative long-term effects on integration of custodial compared to "alternative" sanctions, or of any other types of programmes. Such data are currently being used in two randomised experiments in Switzerland.[13] They are also an attractive alternative to data collected through interviews, given the obvious difficulties in locating and motivating subjects many years after their correctional experience. As the few available examples of long-term evaluations show, one important advantage of controlled trials is precisely the possibility to consider later outcomes in remote areas that no one had anticipated to be causally related to the intervention at stake.

Placebo effects and double-blind trials

To the extent that, in randomised controlled trials, lower re-offending rates have been observed after "alternative" compared to custodial sanctions, or after any sort of treatment, it should not be ruled out that this outcome could be the result of a placebo effect. Indeed, persons sentenced to a custodial sanction who get the "chance" to serve it under the form of an "alternative", i.e., usually without having to go to prison, or who are placed in a "special" treatment unit within confinement,

are offered, in some way, a second (often unexpected) chance which, in turn, may favourably affect their attitudes (as observed by Killias et al. 2000a).[14] A placebo effect has also been envisaged by the authors of the ISSP trial in England, where the experimental group showed moderately lower overall re-offending rates (Little et al. 2004). Subjects in an experimental group may, whenever treatment means avoiding prison or any other unwelcome experience, typically develop the feeling of having been treated "better than expected" or, simply, with more "fairness". Given recent developments in psychological research on "fairness" (Fehr and Rockenbach 2003), such an assessment of the criminal justice system's response is likely to produce, as a result, lower re-offending rates.

In the medical field the obvious answer would be to organise double-blind trials. During the early steps of the Swiss heroin-prescription trial, limited sub-samples of addicts were treated, in a randomised double-blind trial, with various alternative substances (other than heroin). The subjects rapidly discovered, however, that they had not been injected with their "favourite" drug, so that the trial was double blind during a few days at best and had to be terminated (Uchtenhagen 1997). For obvious reasons, double-blind experiments are even less feasible in the field of criminal justice. Any such effects have, so far, found very little attention in the criminal justice literature, including the NRC report where several other threats to internal validity of randomised experiments are discussed (p. 38).

Concluding remarks: the role of evaluations in policy making

If—as often observed in randomised controlled trials—rates of re-offending are similar no matter what intervention the subjects were assigned to, many think the money invested in the experiment has been wasted. Such a view is inappropriate, since experiments are not carried out to show that certain interventions *are* effective. It would not be justifiable to view such results as evidence that "nothing works."

Similar outcomes after carefully evaluated interventions allow researchers and policy makers to conclude validly that the effects of all the options compared are similar. For policy makers, such an outcome means that the choice between programmes can be made on the grounds of consideration beyond their effectiveness, such as relative costs and availability of resources. Further legitimate concerns include fairness and equity to offenders and victims, consistency in sentencing, and popularity among defendants and the public. In the end, criminal law and procedure are searching for equity, not therapy, and corrections and sentences should not be based on treatment considerations so long as there is no evidence of beneficial or damaging collateral effects.

In this sense, striving for evidence-based crime prevention policies does not necessarily restrain policy makers' options as advocated by the Campbell Crime and Justice Group (Welsh and Farrington 2005). It simply means that they will know better what they can expect in adopting solutions or policies whose effects have been documented through meta-analyses that accumulate the evidence from carefully designed experiments. The NRC report certainly comes as a most

valuable source of inspiration to all those who strive to make policy more rational not just through improving meta-analyses but also by bringing more quality into evaluations that form the base of all we know about outcomes.

Notes

1 A good illustration is offered by the impressive review of the literature on the effects of custodial sentences by Smith et al. (2002). In our own review (Villettaz et al. 2006) we found six strong studies (among which two were experiments) conducted in Europe that were not included in the Smith et al. review.

2 For scholars from outside Europe, it may be hard to imagine how little attention Europeans tend to pay to empirical research conducted in other (even neighbouring) countries. Europe's linguistic fragmentation is not the only cause, since ignorance is the rule, even where studies have been published in English or where two or more countries share the same language.

3 In Switzerland, electronic monitoring is being evaluated and compared with community work by two different institutes. The randomised controlled trial (Villettaz and Killias 2005) costs substantially less than a parallel quasi-experimental evaluation.

4 The Swiss heroin trials offer illustrations for both extremes. On one hand, subjects treated with heroin reduced offending by more than 50% and up to nearly 100% (for some offences), i.e., to an extent that exceeded, by far, what usually can be observed after drug treatment (Killias et al. 2005). On the other hand, some of the substances prescribed (as an alternative to heroin) in some randomised sub-trials produced very strong undesirable physical side effects, requiring immediate cessation of those trials (Uchtenhagen 1997). Therefore, and with few exceptions (such as the Geneva experiment by Perneger et al. 1998), heroin prescription was evaluated through a simple before-and-after design.

5 Checking eligibility may also be necessary for some non-custodial sanctions, such as community service and electronic monitoring. Next to the ability and reliability of the convicted defendant to participate in the feasible work assignments, only volunteers can have community service imposed on them, given the European Convention of Human Rights' (Section 10) ban on forced labour.

6 As it turned out in the Swiss community service experiment, subjects admitted by staff without randomisation were living in more difficult conditions than the average of participants. If social workers admitted them because of a particularly positive prognosis, the experience showed that their predictions were worse than average, however.

7 Indeed, community service has been introduced as an arrangement of serving custodial sentences. After the introduction of community service, judges became less reluctant about imposing immediate rather than suspended custodial sentences, knowing defendants had a chance of avoiding prison also in that case. At the same time, judges increasingly imposed sentences not

exceeding 30 days (the maximum that could be served under the form of community service) rather than longer sentences (Killias et al. 2000b).

8 In the present example the Swiss Federal Supreme Court held that the first regulation (introduced in 1976) was unconstitutional (in 1977), pushing the legislator to enact a new law (in 1982) to make the wearing of safety belts compulsory again. Fortunately, safety belt use was monitored over the entire period of repeated change in legislation.

9 This is the overall result. For those convicted of theft (n = 202), the re-conviction rates (after 6 years) were 68% for those sent to prison, compared with 65% for those whose sentence had been commuted. For traffic offenders (n = 1,397), the re-conviction rate was 40% in both groups. Among violent offenders (n = 321), those sent to prison had significantly higher reconviction rates (63% vs 53%).

10 Since prevalence as well as incidence rates fell, after the onset of heroin prescription, to less than 50% (for some offences, even to less than 10%) of pre-intervention levels, distinguishing incidence rate was felt to be less urgent after the first 2 years of the intervention. The long-term evaluation confirmed that crime remained stable at low post-intervention levels.

11 In the case of the community service vs prison experiment (Killias et al. 2000a), only ten (out of 39) participants in the control and 18 (out of 84) in the experimental group had been re-convicted during the follow-up period of 2 years. Thus, statistical power has been critically reduced by the low prevalence of re-convictions in both groups.

12 This was the case, for example, in the Geneva heroin-prescription randomised trial, where the treatment group had higher pre-intervention offending rates than the control group (Killias et al. 2002). Similarly, the experimental group had higher pre-intervention offending rates in the Kent Intensive Supervision and Support Programme (ISSP) trial (Little et al. 2004).

13 One trial (ongoing) compares community service with electronic monitoring (Villettaz and Killias 2005); the other is an update (10 years later) of the Killias et al. (2000a) experiment comparing custodial sanctions with community service.

14 As acknowledged in that publication, Dr. F. Vitaro (University of Montréal) had drawn our attention to this possibility.

References

Aebi, M. F. (2006). *Comment mesurer la délinquance?* Paris: Armand Colin.

Berk, R. A. (2005). Randomized experiments as the bronze standard. *Journal of Experimental Criminology* 1/4, 417–433.

Brennan, P. A. & Mednick, S. A. (1994). Learning theory approach to the deterrence of criminal recidivism. *Journal of Abnormal Psychology* 103/3, 430–440.

Cornish, D. (1987). Evaluating residential treatments for delinquents: A cautionary tale. In K. Hurrelmann, F.-X. Kaufmann, & F. Lösel (Eds.), *Social intervention: Potential and constraints* (pp. 333–345). New York/Berlin: de Gruyter.

Eisner, M. & Ribeaud, D. (2005) A randomised field experiment to prevent violence. The Zurich Intervention and Prevention Project at Schools (ZIPPS). *European Journal of Crime, Criminal Law and Criminal Justice* 13(1), 27–43..

Eisner, M. & Ribeaud, D. (2006) Doing criminological research in culturally diverse contexts. Lessons learned from the Zurich study on the social development of children (in press). *European Journal of Criminology*.

Empey, L. T. & Lubeck, St. G. (1971). *The silverlake experiment*. Chicago: Aldine.

Erez, E. (1985). Random assignment. The least fair of them all: Prisoners' attitudes towards various criteria of selection. *Criminology* 23/2, 365–379.

Fehr, E. & Rockenbach, B. (2003) Detrimental effects of sanctions on human altruism. *Nature* 422, 137–140 (13 March).

Israel, M. & Chui, W. H. (2006). If 'something works' is the answer, what is the question? Supporting pluralist evaluation in community corrections in the United Kingdom. *European Journal of Criminology* 3/2, 181–200.

Killias, M. (1985). La ceinture de sécurité: une etude sur l'effet des lois et des sanction. *Déviance et société* 9/1, 31–46.

Killias, M., Aebi, M. F. & Ribeaud, D. (2000a). Does community service rehabilitate better than short-term imprisonment? Results of a controlled experiment. *The Howard Journal of Criminal Justice* 39/1, 40–57.

Killias, M., Camathias, P. & Stump, B. (2000b). Alternativsanktionen und der ‚net-widening'-effekt. Ein quasi-experimenteller test. *Zeitschrift für die gesamte Strafrechts-wissenschaft* 112/3, 637–652.

Killias, M., Aebi, M. F., Ribeaud, D. & Rabasa, J. (2002). *Schlussbericht zu den Auswirkungen der Verschreibung von Betäubungsmitteln auf die Delinquenz von Drogenabhängigen*. Lausanne: School of Forensic Science and Criminology.

Killias, M., Aebi, M. F. & Ribeaud, D. (2005). Key findings concerning the effect of heroin prescription on crime. In Swiss Federal Office of Public Health (Ed.), *Heroin-assisted treatment. Work in progress*. Berne (Switzerland): Huber.

König, J. M. (2002). *Der Beitrag der Methadonsubstitution zur kommunalen Kriminalprävention. Eine Delinquenzmessung bei Methadonpatienten in Bonn*. Mönchengladbach: Forum Verlag Godesberg.

Lamb, R. R. & Goertzel, V. (1974). Ellsworth house: A community alternative to jail. *American Journal of Psychiatry* 131/1, 64–68.

Legge, I. & Bathsteen, M. (2000). *Einfluss des Methadonprogramms auf die Delinquenzentwicklung polizeibekannter Drogenkonsument/-innen*. Hamburg: Landeskriminalamt.

Little, M., Kogan, J., Bullock, R. & Van der Laan, P. (2004). ISSP: An experiment in multi-systemic responses in persistent young offenders known to children's services. *British Journal of Criminology* 44/2, 225–240.

Lösel, F. & Schmucker, M. (2005). The effectiveness of treatment for sexual offenders: A comprehensive meta-analysis. *Journal of Experimental Criminology* 1/1, 117–146.

Lösel, F., Beelmann, A., Stemmler, M. & Jaursch, S. (2006). Prävention von Problemen des Sozialverhaltens im Vorschulalter. *Zeitschrift für Klinische Psychologie und Psychotherapie* 35/2, 127–139.

Mannheim, H. & Wilkins, L. T. (1955). *Prediction methods in relation to Borstal training*. London: HMSO.

McCord, J. (1990). Crime in moral and social contexts. *Criminology* 28/1, 1–26.

National Research Council of the National Academies (NRC). (2005). *Improving evaluation of anticrime programs*. Washington DC: The National Academies Press.

Olds, D. L. & Kitzman, H. (1993). Review of research of home visiting for pregnant women and parents of young children. *The Future of Children* 3/3, 53–92.

Ortmann, R. (1994). Zur Evaluation der Sozialtherapie. Ergebnisse einer experimentellen Längsschnittstudie zu Justizvollzugsanstalten des Landes Nordrhein-Westfalen. *Zeitschrift für die gesamte Strafrechtswissenschaft* 106/4, 782–821.

Ortmann, R. (2000). The effectiveness of social therapy in prison. A randomised experiment. *Crime and Delinquency* 46/2, 214–232.

Perneger, T. V., Giner, F., Del Rio, M. & Mino, A. (1998). Randomised trial of heroin maintenance programme for addicts who fail in conventional drug treatment. *British Medical Journal* 3/17, 13–18.

Petrosino, A. & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism. Results from meta-analyses of experimental and quasi-experimental research. *Journal of Experimental Criminology* 1/4, 435–450.

Ribeaud, D. (2004). Long-term impacts of the Swiss heroin prescription trials on crime and treated heroin users. *Journal of Drug Issues* 34/1, 163–194.

Schweinhart, L. J., Barnes, H. V. & Weikart, D. P. (1993). *Significant benefits: The high/scope Perry Preschool study through age 27*. Ypsilanti (MI): High/Scope Press.

Sherman, L. & Weisburd, D. (1995). General deterrent effects of police patrol in crime hot spots: A randomized experiment. *Justice Quarterly* 12/4.

Smith, P., Goggin, C. & Gendreau, P. (2002). *Effets de l'incarcération et des sanctions intermédiaires sur la récidive: Effets généraux et différences individuelles*. Ottawa: Solicitor General of Canada.

Tilly, N. (2000). Experimentation and criminal justice policies in the United Kingdom. *Crime and Delinquency* 46/2, 194–213.

Uchtenhagen, A. (1997). *Versuche für eine ärztliche Verschreibung von Betäubungsmitteln. Synthesebericht*. Zurich: Institute for Addiction Research.

Van der Werff, N. (1979). *Speziale preventie*. The Hague: WODC.

Van der Werff, N. (1981). Recidivism and spezial deterrence. *British Journal of Criminology* 21/2, 136–147.

Villettaz, P. & Killias, M. (2005). *Les arrêts domiciliaires sous surveillance électronique: Une sanction ≪expérimentale ≫*. Lausanne: Institute of Criminology and Penal Law of the University of Lausanne.

Villettaz, P., Killias, M. & Zoder, I. (2006). *The effects of custodial vs. non-custodial sanctions on re-offending. A systematic review of the state of knowledge*. Campbell Collaboration Crime and Justice Group (under review).

Weisburd, D. (2000). Randomized experiments in criminal justice: Prospects and problems. *Crime and Delinquency* 46/2, 181–193.

Weisburd, D., Petrosino, A. & Mason, G. (1993). Design sensitivity in criminal justice experiments. *Crime and Justice* 17, 337–379.

Welsh, B. C. & Farrington, D. P. (2005). Evidence-based crime prevention: Conclusions and directions for a safer society. *Canadian Journal of Criminology* 47/2, 337–354.

Wilkins, L. T. (1969). *Evaluation of penal measures*. New York: Random House.

About the author

Martin Killias has been Director of the Institute of Criminology and Criminal Law of the University of Lausanne (Switzerland) and is currently Professor of Criminology and Criminal Law at the University of Zurich (Switzerland). He has degrees in Sociology/Social Psychology and Law. He has been a postdoctoral fellow at the School of Criminal Justice at the University of Albany (NY) and visiting professor in Canada, the United States, Italy, England and the Netherlands. He has directed several randomized controlled experiments in the field of corrections in Switzerland. In 2001-2002, he served the European Society of Criminology as its first president.