Journal of Experimental Criminology (2021) 17:151–160
https://doi.org/10.1007/s11292-019-09406-w

External validity and assignment of experimental vs. control treatment providers within small work groups: a research note

Lawrence W. Sherman[1,2] · Sarah van Mastrigt[3] · Christian B. N. Gade[3] · Theresa Ammann[3] · Heather Strang[1]

Published online: 22 January 2020

Abstract

Objectives When offenders or victims are randomly assigned to receive experimental vs. current treatments, the external validity of results may depend on whether different *treatments* are delivered by similar kinds of treatment *providers*. When treatment providers volunteer to deliver innovative practices in an experiment, it is unclear whether outcomes depend on the content of the treatment, enthusiasm of the providers for the new practice, or both. In such situations, the potential for what we describe as *differential predisposition* of volunteers for a new treatment raises a question of external validity.

Methods We describe the process by which 14 out of 29 mediators across seven Danish police districts came to deliver a new, restorative conferencing method of conducting face-to-face meetings between offenders and their victims, in comparison to longstanding mediation methods.

Results We negotiated with all seven District mediation leaders and all 29 of their mediators to use partial random assignment of 14 of the mediators to deliver the new, restorative model. The 14 trained providers of the new method were substantially similar in several measureable characteristics to the 15 other mediators who continued to use the preexisting model, but we cannot measure directly the extent or balance of their predispositions for delivering each model.

Conclusions While small work teams pose obstacles to simple random assignment of *treatment providers* to deliver experimental practices, the random assignment of *victims and offenders* to two different models of service might be made more externally valid by use of partial random assignment of service providers.

Keywords External validity · Differential predisposition · Partial random assignment · Treatment providers · Small work groups · Restorative justice conferences · Mediation

✉ Lawrence W. Sherman
Lawrence.Sherman@crim.cam.ac.uk

Extended author information available on the last page of the article

Introduction

Differential predisposition among service providers testing an innovation

The primary aim of conducting randomized experiments in criminal justice is to develop a generalizable body of knowledge about what works to reduce crime (Lipsey et al. 2006). Yet, as the "replication crisis" (Losel 2017) in social science suggests, the advances of experimental criminology have focused more on internal validity of each test than on the external validity of test results when applied to other populations.

One key issue in the external validity of experiments in providing services to victims and offenders is the characteristics of *people providing the services*. Yet, these characteristics are rarely made explicit or reported in detail in peer-reviewed journal articles. More importantly, the predispositions of service providers to carry out the new service are not explicated in relation to the new treatment being provided, at least not at the individual level.

Our central concern in this research note is that when small numbers of people deliver a new treatment, even to a large and randomly assigned sample, the extent to which those people are predisposed to perform that treatment can be a critical dimension of external validity of the results. If the providers' predisposition to a new model is strong that fact may lead to effects which cannot be replicated when providers of lesser predisposition deploy the same treatment. Conversely, if the initial test shows weak effects, it remains possible that stronger effects could be obtained by service providers whose personalities and values make them more predisposed to carrying out the new model.

In training police to use procedural justice principles in citizen interactions, for example, one randomized controlled trial (RCT) was conducted in a context of a "training academy environment that emphasized aggressive policing and officer safety" (Rosenbaum and Lawrence 2017: 293). The authors duly—and unusually—noted that this environment was not predisposed to support a new model of how to train police for public contact. That model, and the characteristics desired for the few instructors who would teach it to a large sample, was clearly described as follows:

> Instructors were trained in the new curriculum. Part of the instructor training focused on creating a culture where integrity of the message, using the right tone and appropriate pedagogy, is critically important. Instructors were sensitized to the fact that they can sometimes send mixed messages to the recruits, and, consequently, undermine their own training objectives. Instructors were also trained in the professionalism of teaching and the importance of role-modeling, including starting and ending class on time, treating every student with dignity and respect, evaluating students using fair and objective standards, etc. (Rosenbaum and Lawrence 2017: 304)

Quite apart from the content of procedural justice (PJ) training, the above-quoted message in training the trainers is that the PJ training is very much about *precision:*

precision in talking to people, in starting class on time, in using objective standards, etc. Yet, people vary widely in their predispositions for using precision. Even among physical therapy graduates, for example, there is wide variation in a predisposition for or against seeking precision (Rovezzi-Carroll and Leavitt 1984). There are many values that may be affected by any new method of delivering services, and people may vary widely on those values—or any other values embedded in the innovation they are asked to embrace and carry out themselves.

So when an experiment tests an innovation, something more important than the work role or demographics of those delivering the innovation may be at stake. That something is what we call *differential predisposition*, as a short-hand label for a range of characteristics that may flow out of providers' personalities that makes them more or less suitable for carrying out a new idea. In our experiment in Denmark, the predispositions included, for example, qualities stressed in our own training like:

- displaying "a calm character"
- listening without interrupting people
- asking Socratic questions rather than proposing specific actions

or other elements of how the treatment is delivered through interactions with victims and offenders. While those specific characteristics are examples drawn from the experimental treatment that is the subject of this research note, they illustrate the kinds of qualitative dimensions that could be correlated with either volunteering for delivering a new treatment or being picked by supervisors as being a "good fit" for that particular experimental treatment.

The potential selection bias of treatment providers as a threat to external validity, then, is not just about enthusiasm or willingness to comply with a new way of interacting with offenders and victims. It is about whether their personalities make them more or less predisposed to "get" and "grasp" the application of the particular innovation in question. This predisposition is multidimensional, with different aspects of a personality making the individual more or less likely to offer a "match" to the requirements of each innovation. This matching may occur in ways that might—or might not—be replicated by all treatment providers in an effectiveness trial. Thus, it is a potential issue in a general scaling-up to a rollout—using wider implementation of the innovation (Gottfredson et al. 2015).

The effect sizes for any rollout will depend not only on provider predisposition to the innovative treatment but also on predispositions to the counterfactual treatment provided by control group providers. Even if the control group providers are enthusiastic about the control treatment, they might become very unenthusiastic if asked to change over to the innovation. Such resistance to the newly tested innovation could, in any number of ways, make the effectiveness of the innovation much weaker than it was in an initial test delivered by enthusiastic, purposively selected volunteers.

Quite apart from the preferences of treatment providers, issues of generalizability remain. Even if all providers in a scaling-up were cooperative and did the best they could, they might indeed not be *as good* a fit for the experimental treatment as people who self-selected or who were selected by supervisors for the initial pilot test (efficacy trial) of the innovation (Gottfredson et al. 2015).

As far as we can tell, this is not an issue that has previously been raised in the methodological literature on randomized field experiments. We should acknowledge that premise and offer for discussion and falsification the hypothesis that there is such a dimension as differential predisposition to particular tactics for dealing with crime victims and offenders. We therefore define differential predisposition for an innovation as *the degree to which the experimental service providers manifest behavior that makes them an atypically good match for delivering the experimental program*. This may include their intensity of belief in or commitment to the program's theory, its morality, and promise of success. It may also include observable dimensions of their interactions, such as measures drawn from police video clips that have been used to assess providers' delivery of the dimensions of procedural justice (Nawaz and Tankebe 2018). All of this should be highly correlated with the integrity with which each human service provider delivers the model of interaction as trained to do.

The difference in predisposition of providers to deliver an innovative model with integrity may explain, for example, the well-documented "developer effect" (Petrosino and Soydan 2005): the difference between findings of evaluations of programs led by a developer of the program showing large benefits, and weaker or no benefits found in evaluations of the same program conducted by evaluators who had not been developers of the program. Eisner (2009) suggests this effect may be caused by developers' conflicts of interest when they want to sell a commercial treatment product, leading to systematic biases in conducting the evaluation. It seems equally plausible, however, to see the difference as higher integrity of program implementation with developers than with non-developers (Sherman and Strang 2009). If developers are given freedom to pick the "best" people for the innovation, they may get better results because of a differential predisposition.

Where evaluators are given power to recommend how a sample of service providers should be selected to deliver an innovation, there is a strong argument that they should always select the best people available—often by calling for volunteers. In many experiments (Sherman and Berk 1984; Sherman et al. 1991; Strang et al. 2013), for example, the only way the experiment can be conducted at all is to call for and persuade volunteers to deliver the experimental practice. These volunteers may give the program a high-integrity delivery that yields strong internal validity of the test. Yet, because of the way in which they were selected, the result may lack external validity for predicting results of a scaled-up policy that requires all personnel to deliver an innovation (regardless of their predisposition for that policy).

A partial random assignment solution

A special variety of this challenge is in small service provider work groups (e.g., under ten people per group). In such experiments, the design may require a large proportion of the group to participate in the experiment. Under those conditions, the workers' opinions about the innovation itself may be divided. The default option is to ask for volunteers to deliver the innovation, which by definition may attract people who are favorably disposed to the new practice. Thus, clients (such as victims or offenders) are randomly assigned to be dealt with by volunteers for the new method vs. providers of the current method who chose not to change. Alternatively, all (or most) of the workers

may wish to try the new method, which would make random assignment to the role of providing the innovation an ideal solution.

More common, however, may be the middle case: the group shows a wide general interest in the innovation, but a collection of special circumstances block some people from taking it up. Under those conditions, some might default to an all-volunteer assignment. There is a strong case, however, for another method to provide great external validity: *partial random assignment*. To the extent that a larger segment of the work group may be willing to accept random assignment into the experimental program, there is a theoretical benefit in equalizing predispositions to the methods of both the innovation and the previous business as usual model.

This note describes how the authors implemented the strategy of partial random assignment in launching an experiment with the Danish Police in 2017–2018.

Setting: random assignment of mediation vs. restorative justice conferences

Since 2010, Danish law has mandated all police areas to provide opportunities for crime victims and their admitted offenders to meet to discuss their cases. These meetings are typically based on a "triangle" of offender(s), victim(s), and a mediator. They occur not as a diversion from prosecution but as an add-on to criminal justice at almost any stage in the process, including offenders serving prison sentences. Each of Denmark's 12 current police districts has a police employee serving as the main coordinator of the offers made to offenders and victims to attend such meetings, which are led by non-police who are trained as mediators (Gade 2018; Storgaard 2015).

In 2017, the Danish Police agreed to conduct an RCT to compare the mediation model ("Konfliktråd" in Danish or what we call "KR" in this research note) to a different form of meetings between victims and offenders called restorative justice conferences or "RJCs" (Sherman et al. 2013; Sherman et al. 2015a). The RJC structure, compared to the KR method, typically includes more participants than KR, a less engaged style for the "convenor" (called mediators for KR and facilitators for RJCs), a longer conversation with more engagement of all participants, and a more structured approach to having offenders leave the meeting with an agreement about repair of victim harm and what offenders will do to help reduce their chances of committing more crime in the future (Kyvsgaard et al. 2018). The agreement was that the RJCs in the experiment would be delivered solely by persons who were already employed by the program as mediators. Those mediators selected for the experimental (RJC) treatment would receive additional training in the experimental method from Australian RJC trainer John McDonald, who had trained most of the facilitators involved in the previous UK and Australian randomized trials of RJCs (Sherman et al. 2015b).

Choosing whether to participate in the experiment was ultimately left to the discretion of the 12 District Police commanders in Denmark, on a district-by-district basis. In total, 7 of the 12 districts elected to participate. Each of those districts had one or more local KR coordinators and a network of mediator consultants who were paid on a case-by-case basis. The number of consultant mediators in each of the seven districts in late 2017 ranged from two to seven. At the time the experimental convenors were selected for the additional training in late 2017, we knew that the total number of mediations (KR) conducted in those districts in 2016 had been 396, and the number of cases

deemed eligible for the experiment was 174, after exclusions for intimate partner violence, traffic cases, and neighbor disputes. The kinds of cases established as eligible for the RCT included both property and violent crimes, excluding murder and rape but including robbery and assaults.

The RCT plan was to enter into the experiment all of the eligible cases to be treated in a meeting of consenting victims and offenders, after both had also granted consent to participate in the experiment. The experiment was then to use random assignment to channel each case to either KR or RJC methods. Thus, roughly half of the eligible cases were to be randomly assigned, by a research team at the Aarhus University, to consultants delivering one or the other of the two treatments. To the extent possible, the plan was to make the experience of KR vs. RJC as different as possible by having differently trained treatment providers for each group of clients.

The question the planning group struggled to answer was how to minimize differential predisposition among the providers of the experimental treatment, while maximizing the quality and integrity of delivery of both methods. This was especially important because the core of the experiment was in providers behaving in very different ways: speaking less in the innovative treatment (vs. more in the established method), being less (vs. more) directive, and in hearing from more (rather than fewer) other people in the room who can discuss their emotions and hopes for the future.

The planning group was fortunate to include an active mediator with several years of experience. This mediator was attuned to the practical challenges the coordinators faced in obtaining cases, consent by participants, and convenors for the meetings. This knowledge combined with that of other members of the planning group trained in experimental design. Issues of external validity and various options regarding the selection of service providers to be trained in the new method were first discussed with each district coordinator separately, followed by a national consultation meeting including all coordinators and service providers from the seven participating police districts.

The key principle sought for the experimental design was to minimize the extent of switching back and forth between one method and the other by each individual convenor. Experimentalists in the planning group recommended that no cases in the randomized trial be conducted by convenors who use both methods with eligible cases. In order to develop the skill and consistency of RJC facilitation, the planning team recommended that for the duration of the experiment, all of the cases given to convenors trained as RJC facilitators would use the RJC method. The coordinators agreed to this, with the exception of allowing some ineligible cases to be assigned to KR methods delivered by the RJC-trained facilitators. This was one of many necessary compromises in structuring the workload of such small service provider groups.

Issues in selecting KR mediators for RJC training

The issues in selecting KR mediators for RJC training were driven more by logistical considerations than by the personal preferences of the convenors. To be sure, some mediators simply did not want to try the new method, so they declined to be randomly assigned to receive the training or not. Others, however, simply had time and availability problems. Since most of the mediators had other jobs as their primary employment, some were simply unavailable to attend a 3-day training session. Others were

new mediators, who wanted to gain skill in KR before being trained in alternative methods.

Based on the abovementioned practical challenges, the coordinators in some districts had to nominate some or all of their remaining mediators to switch to the RJC method. A total of six out of the 14 mediators trained in RJC were assigned to the training using this nonrandom and non-self-selected, method of assignment.

A further consideration was a demographic balance by gender, as well as by age. Variation in months, and even years, of experience in mediating by the KR method was also a concern. In the districts where the number of service providers available to be trained in the new method exceeded the number of training spots available, random assignment to RJC training was thus employed. Eight of the 14 individuals ultimately trained in the RJC method were randomly assigned.

Results

In the course of the national meeting, the *partial* random assignment was conducted from the purposively selected list of those for whom coordinators and mediators had agreed would be randomly assigned. The process was transparently explained in a plenary session and was demonstrated to the entire group as an example of how random assignment would work in the experiment. At the same time, the plenary group knew that some members had declined random assignment, and other had been assigned to training (not under protest; just not as volunteers).

Table 1 shows the result of this partial random assignment. With a limited selection of characteristics of the 29 mediators participating in the RCT, the two groups are reasonably similar. To the extent that differential predisposition is a particular threat to external validity (by potentially over-estimating the effect of RJCs), it is worth noting that fully 57% of the RJC facilitators were selected by random assignment. This may or may not have reduced their predisposition for the innovation, but at least some of them had clearly been uncertain about whether they wanted to take the training and change methods. For some of the 57%, the decision to adopt the experimental method was far from entirely self-directed.

At the same time, the balance of 43% who were assigned to the RJC training by coordinators may have been similarly equipoised as the randomly assigned facilitators, with their position chosen by the coordinators—and their own decision not to object (or quit).

Table 1 Differences in selected treatment provider characteristics

Selected characteristics $N = 29$	KR mediators, $N = 15$	RJC facilitators, $N = 14$	Relative % difference RJC to KR
Randomized into task (% yes)	40.0	57.1	+ 42%
Gender (% female)	73.3	50.0	− 32%
Age (years)	53.6	52.1	− 3%
Konfliktråd experience (months)	45.3	67.7	+ 49%

As for the balance between the RJC and KR treatment providers, the small number of people (29) and even smaller number of those randomly assigned made it unlikely that the balance would be very precise. Yet, the differences between the two groups are not large. Table 1 shows that the largest difference was a relative difference of 49% higher experience at KR mediation in the RJC-assigned group compared to the KR group. The KR group was also 46% more female than the RJCs. The mean age of the consultants in the two groups was almost identical.

Discussion

These results are far from definitive. Most importantly, they do not directly measure our key concept of "differential predisposition." In theory, we might have done so, perhaps by interviewing each service provider about their comfort level with each dimension of the innovation (such as talking less than they have in the previous several years of meeting with victims and offenders). Given the small sample, we could not block on such answers and randomly assign within different response groups. Instead, we focused on the quest for maximizing random assignment as having the best chance to minimize *any* differences between treatment and control providers that might have affected the outcomes of the respective treatment groups.

What Table 1 does not show is any measure of personality, or even elements of practices that convenors employ when chairing a meeting between crime victims and offenders. In the design of the experiment, however, observers in a sample of cases in both treatment arms will record the elements of behavior the convenors demonstrate in these meetings. To the extent that there is low variance within each group across a range of those behaviors (such as telling all participants what the procedures will be at the outset of the meetings) that will strengthen the internal validity and even power of the test.

The results may be useful in illustrating the extent to which even discussing external validity as an element of experimental design may be important to consider. Our aim was to find the "least worst" solution to the challenge of making the providers of two different treatments as similar as possible. The result may be greater causal identification of any differences in outcomes as associated with the treatment content, rather than the characteristics of the treatment providers. We cannot know for sure, but we contend that this approach is more justified than simply using volunteers or letting supervisors select the most predisposed to each treatment model.

External validity of the experimental results would be most enhanced by replications, as well as larger numbers of consultants providing the treatments. Yet in order to replicate an experiment, scientists need to know as much about a completed experiment as possible. Experimental criminology has long been afflicted by a large "black box" in the middle of the design, in which issues such as the selection of key personnel have been left unaddressed. Perhaps the major value of this research note is to demonstrate the kind of documentation that can help to make experiments more replicable.

As Winchester's (2018) recent history of precision engineering indicates, the long arc of scientific history has moved from lesser to greater precision. In the development of experimental criminology, there is much to be gained by more precision when it is possible to achieve it. Even when it is not possible, it may be useful to have such

research notes as this one for "next time," offering a framework for what might be a preferred research design.

Acknowledgments Special thanks go to the mediators and staff of the Danish Victim-offender Mediation Programme for their ongoing co-operation and dedication to the Konfliktråd Impact Project. Thanks also go to the two anonymous reviewers for their thoughtful comments on an earlier draft of this paper.

Funding information The authors would like to thank the Danish foundation TrygFonden for funding this research.

References

Eisner, M. (2009). No effects in independent prevention trials: can we reject the cynical view? *Journal of Experimental Criminology, 5*(2), 163–183.

Gade, C. B. N. (2018). "Restorative justice": history of the term's international and Danish use. In A. Nylund, K. Ervasti, & L. Adrian (Eds.), *Nordic mediation research*. Cham: Springer.

Gottfredson, D., Cook, T. D., Gardner, F. E. M., Gorman-Smith, D., Howe, G. W., Sandler, I. W., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: next generation. *Prevention Science, 16*, 893–926. https://doi.org/10.1007/s11121-015-0555-x.

Kyvsgaard, B., van Mastrigt, S., & Gade, C. B. N. (2018). Genoprettende retfærdighed og recidiv i Danmark. *Samfundsøkonomen, 4*, 23–28.

Lewin, K. (1946). *Force field analysis*. *The 1973 annual handbook for group facilitators*, 111–113.

Lipsey, M., Petrie, C., Weisburd, D., & Gottfredson, D. (2006). Improving evaluation of anti-crime programs: Summary of a National Research Council report★. *Journal of Experimental Criminology, 2*(3), 271–307.

Lösel, F. (2017). Evidence comes by replication, but needs differentiation: the reproducibility issue in science and its relevance for criminology. *J Exp Criminol.* **https://doi.org/10.1007/s11292-017-9297-z**. Online publication, (18 August).

Nawaz, A., & Tankebe, J. (2018). Tracking procedural justice in stop and search encounters: coding evidence from body-worn video cameras. *Cambridge Journal of Evidence-Based Policing, 2*, 139–163.

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.

Rosenbaum, D. P., & Lawrence, D. S. (2017). Teaching procedural justice and communication skills during police–community encounters: results of a randomized control trial with police recruits. *Journal of Experimental Criminology, 13*, 293–319.

Rovezzi-Carroll, S., & Leavitt, R. (1984). Personality characteristics and expressed career choice of graduating physical therapy students. *Physical Therapy, 64*, 1549–1552.

Sherman, L. W., & Strang, H. (2009). Testing for analysts' bias in crime prevention experiments: can we accept Eisner's one-tailed test? *Journal of Experimental Criminology, 5*(2), 185–200.

Sherman, L. W., & Berk, R. A. (1984). The specific deterrent effects of arrest for domestic assault. *American Sociological Review 49*, 261–272.

Sherman, L. W., Schmidt, J. D., Rogan, D. P., Gartin, P. R., Cohn, E. G., Collins, D. J., & Bacich, A. R. (1991). From initial deterrence too longterm escalation: shortcustody arrest for poverty ghetto domestic violence. *Criminology 29*(4), 821–850.

Sherman, L. W., Strang, H., Barnes, G., Woods, D. J., Bennett, S., Inkpen, N., & Slothower, M. (2015a). Twelve experiments in restorative justice: the Jerry lee program of randomized trials of restorative justice conferences. *Journal of Experimental Criminology, 11*(4), 501–540.

Sherman, L. W., Strang, H., Mayo-Wilson, E., Woods, D. J., & Ariel, B. (2015b). Are restorative justice conferences effective in reducing repeat offending? Findings from a Campbell systematic review. *Journal of Quantitative Criminology, 31*(1), 1–24.

Storgaard, A. (2015). Denmark. In F. Dünkel et al. (Eds.), *Restorative justice and mediation in penal matters: a stock-taking of legal issues, implementation strategies and outcomes in 36 European countries*. Forum Verlag Godesberg: Mönchengladbach.

Strang, H., Sherman, L. W., Mayo-Wilson, E., Woods, D., & Ariel, B. (2013). Restorative justice conferencing (RJC) using face-to-face meetings of offenders and victims: Effects on offender recidivism and victim satisfaction. A systematic review. *Campbell Systematic Reviews 9*(1), 1–59.

Winchester, S. (2018). *The perfectionists: how precision engineers created the modern world*. NY: HarperCollins.

Affiliations

Lawrence W. Sherman[1,2] · Sarah van Mastrigt[3] · Christian B. N. Gade[3] · Theresa Ammann[3] · Heather Strang[1]

[1] University of Cambridge, Cambridge, UK

[2] University of Maryland, Maryland, USA

[3] Aarhus University, Aarhus, Denmark