University of Groningen

Doing field studies. What is it all about?

Keizer, Kees; Lindenberg, Siegwart; Steg, Linda

Published in:
Group Processes & Intergroup Relations

DOI:
[10.1177/1368430213510750](https://doi.org/10.1177/1368430213510750)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2014

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Keizer, K., Lindenberg, S., & Steg, L. (2014). Doing field studies. What is it all about? *Group Processes & Intergroup Relations*, 17(3), 404-410. <https://doi.org/10.1177/1368430213510750>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Doing field studies. What is it all about?

Group Processes & Intergroup Relations

2014, Vol. 17(3) 404–410

© The Author(s) 2013

Reprints and permissions:

sagepub.co.uk/journalsPermissions.nav

DOI: 10.1177/1368430213510750

gpir.sagepub.com

Kees Keizer, Siegwart Lindenberg, and Linda Steg

Abstract

Field studies provide much ecological validity but are limited with regard to randomization and control. What can be and what is done about it? Because so few psychologists do field experiments, little is known about it. In our reply to Wicherts and Bakker (in press), we show what goes on behind the scene in doing field research, and we discuss the importance of cumulative results by *inclusive replications* that cover strict replications, but also replications in different settings, using different paradigms.

Keywords

field experiments, littering behavior, methodology, norms, validity

Paper received 30 July 2013; revised version accepted 12 September 2013

Introduction

Somebody once compared doing field experiments to washing dishes. Even though the water is somewhat murky, dishes come out cleaner. In other words, field studies have some weaknesses (compared to lab experiments) but also some considerable strengths. The strengths lie in the ecological validity (see also Cialdini, 2009) and the weaknesses in the difficulty of randomization and controlling all sorts of factors that might be of influence. In the article “Broken Windows, Mediocre Methods, and Substandard Statistics,” Wicherts and Bakker (in press) put forward some methodological concerns with some of our field studies on norm-violating behavior which were reported in Keizer, Lindenberg, and Steg (2008) and Keizer, Lindenberg, and Steg (2011). Even if it was not their intention, the authors invite an interesting discussion on the weakness as well as

the merits of conducting field studies. In this paper we aim to shed some light on the inevitable problems and challenges one will encounter when working in the field and explain how we tried to limit the negative impact of these factors as to secure internal validity of our results.

We are the first to acknowledge that our field studies have some weaknesses, as all (field) studies do. The authors make some valid points. However, their criticism has two strong limitations. First, there is the armchair problem. For

University of Groningen, The Netherlands

Corresponding author:

Kees Keizer, Faculty of Behavioral and Social Sciences,
University of Groningen, Grote Kruisstraat 2/1, 9717 TS,
Groningen, The Netherlands.
Email: K.E.Keizer@rug.nl

those who have not done field experiments, it seems quite mysterious why the researchers did not keep track of everything in the environment that might possibly be of relevance. Thus, the authors' critique of our field experiments seems quite misleading about the possibilities field experimenters have and the adequacy of standard lab methodology for field settings. Second, the authors fail to pay any attention to the cumulative results from our and others' (such as Cialdini, Reno, & Kallgren, 1990) field experiments. True, in every field experiment, there are factors that remain uncontrolled. Yet, our confidence in their conclusions grows as variations in the paradigms used in our field experiment and other studies show convergent results.

For colleagues who are not quite familiar with this branch of research, it is important to know more about what goes on behind the scene when doing field experiments. There is much thought going into limiting the possible handicaps before running the actual field experiment. We will discuss some of the shortcomings mentioned by the authors, as well as some additional ones. We will also take a closer look at the adequacy of the propositions made by the authors to overcome certain weaknesses, and how many of the procedures proposed work very well in the lab, but are likely to fail (as we often found out by trial and error) to meet their objective in the field. We will present alternative procedures we adopted to enhance experimental control in the field and hence internal validity, and explain why we adopted these procedures. Also, we will explain the need for replications and the use of multiple methods as a key strategy to secure scientific rigor of research.

Searching for Cross-Norm Effects

The aim of our original study in 2008 was not testing the broken window theory (Wilson & Kelling, 1982). Rather, we tested the cross-norm hypothesis derived from goal framing theory (Lindenberg & Steg, 2007): observing (signs of) a norm violation by others would increase the

likelihood of violating other norms, because norm violations weaken the relative strength of normative goals and strengthen hedonic and/or gain goals. Yet, our findings have important implications for (the feasibility of) the broken window theory, that predicts that norm transgressions will lead to more severe norm transgressions and even crimes. The broken window theory has been widely adopted by practitioners like policy makers and law enforcers around the world to regulate behavior in the public domain, although sound experimental support was lacking. This relevance for the broken window theory was one of the main reasons why we decided to test our hypotheses in a field setting. Although lab studies undeniably would have given us more experimental control, and (not unimportant) would have saved us a lot of time, it would have severely limited one very important strength of our field approach: a high external validity. Well-designed lab studies are indeed high in internal validity. However, external validity is typically low because of the artificial setting, the lack of natural human activities, and because of the fact that they are commonly conducted among (first year) psychology students. Yet, this price of lab studies is often not explicated. We anticipated that the conclusions of our studies would have important practical implications and would guide practitioners in the field. Therefore, we felt that our research should be understandable and insightful for this audience and ensuring high external validity should be a key point in our studies. This guided two main decisions regarding the experiments: we wanted to measure actual behavior as a dependent variable, and we wanted the studies to take place in relevant public settings. However, let's consider the costs of this decision and how we tried to limit these costs.

Battling Selection Effects

In our first study on the potential spreading effect of norm-violating behavior (Keizer et al., 2008), we studied among other things whether people were more likely to litter a flyer that was attached to the handlebar of their bicycle in a setting with

graffiti than a setting without graffiti. The idea of using littering as the dependent variable was taken from field experiments conducted by Cialdini and his coworkers (e.g., Cialdini, Reno, & Kallgren, 1990) that have become classic by now. As the authors correctly point out, this is a quasi-experiment as we did not randomly assign participants to the conditions. Random assignment was simply not possible in this public setting, as it would have entailed randomly spraying graffiti or painting over the alley's wall (with an exactly similar end result) and removing it over and again. Wicherts and Bakker (in press) suggest we should have followed Reiter and Samuel (1980) because we would have avoided using different days of the week. However, it is exactly the other way around. Knowing Reiter and Samuel (1980), they inspired us to avoid some of their limitations. This is quite instructive and worth relating as it allows a good look behind the scenes of doing field studies. In the littering experiment, to concentrate the study with regard to days, we considered the possibility of having a setting with multiple parking areas, as did Reiter and Samuel (1980), which would entail spraying one area with graffiti and the other not. However, we felt that such an approach would not enhance internal validity (the main reason for using random assignment) for a number of reasons. First, the areas should be highly similar as any differences in place would be possible confounds.

Second, such a design would not rule out the possibility of selection effects, which would be another confound. Often one of the areas was closer for the people looking for a parking space than the other, which would possibly have resulted in systematic differences between participants in the areas. Also the closest area would fill up first thereby taking away the main advantage of such a set-up, simultaneous testing. In fact, such selection effects have probably played a role in the study of Reiter and Samuel (1980). Indeed, it is very likely that the first floors of the parking structure, mainly used by workers of neighboring businesses, were filled up first with those who came to work early. Repeatedly alternating conditions across areas could have been a

solution here, but was not employed by Reiter and Samuel as the sign manipulation was not systematically varied over the different floors, probably because this was not feasible.

Third, the settings (with multiple areas) that seemed most appropriate in our case had various other problems that would have seriously threatened the internal validity of our studies. For example, they were too crowded, with peak moments in which many people collected their bicycle at the same time (where the presence of others may have seriously affected our results), or had a very specific visitors group like students.

To further reduce the risk of self-selection, we decided to attach flyers to bicycles instead of handing people a flyer. Many potential participants will simply refuse to take the flyer, and hence could not have been considered as participating in the study. There was another important advantage of attaching the flyers to the bicycles. As stated, we wanted this specific study to take place in a public setting, where other people are present, but where the number of people present would not influence a participant's perceived chances of getting (socially) sanctioned (as this would be a possible confound). We reduced the likelihood of informal sanctioning because the bicycle of the participant (and the other bicycles present) offered people a shield that allowed them to litter unnoticed by the public. As we explained in our correspondence with the authors of the critique, we decided *before* the start of the experiment not to include people that arrived at the same time as another person who had parked their bicycle right next to them. We reasoned that in such cases concealed littering would have been very difficult for the participants, which would have introduced sanction probability as a confound.

The location of the setting and the time of day at which the experiment was run played an important part in overcoming various lurking methodological problems. As the authors rightfully point out, there is the threat of the same participant being recorded in different conditions or even in the same condition (as was the case in the Reiter and Samuel [1980] study). As explained

next, we abandoned the idea of recording or photographing participants for ethical code reasons. We therefore carefully looked for settings that would not attract daily visits by the same people. Settings near local supermarkets were therefore discarded. The present setting was near shoe shops and clothing shops, where one would not expect frequent visits by the same people. At certain days there is indeed a local market nearby, however visitors of this market seem to park their bicycles alongside the market, which is more convenient. The location was selected also for the fact that participants presented themselves with time intervals between them. Although this lengthened the total duration of the experiments greatly, it reduced the risk of people observing other participants littering or not littering (which would have affected the likelihood that they littered; cf. Cialdini et al., 1990) and reduced the risks of effects of sanction probability, as explained before. Also the layout of the alley (i.e., people facing the wall when picking up their bicycle) and the often concealed littering further reduced the risk of observing others littering or not (but did not completely eliminate it). The other settings of the experiments in the Keizer et al. (2008) studies were carefully selected with similar criteria in mind. In the study on the impact of unreturned shopping carts on littering we did not select a neighborhood supermarket, but we deliberately picked a megastore located outside the city, which is commonly used for (weekly) bulk shopping. The location of the graffiti and litter studies in the 2008 and 2011 papers also had several other advantages over other settings that contributed to making it our setting of choice. For example, the settings allowed out-of-side observation to such an extent that we could guarantee that observer effects would not take place. Also, the locations were typically visited by a cross-section of the population, more so than when we had chosen a specific neighborhood.

The authors note that the differences between conditions found may be the result of testing on different days. They seem to be overly optimistic regarding the time involved conducting such field experiments. It took various weeks to execute

these studies (see our explanation of the laborious procedure followed in the previous lines), and hence every condition included multiple days. At forehand we decided to solely focus on weekdays, as the characteristics of the setting and the visitors are rather different on Saturdays and Sundays compared to weekdays. Indeed, as the authors point out, there are extended shopping hours on Thursdays. However, the experiments were run in the afternoon, so this posed no problem for our experiments.

Choosing the Right Protocol

Another concern expressed by the authors of the critique is that in our studies the observer was not blind to the hypothesis and reason behind the experiments. This is indeed a valid concern. We considered the use of a “blind” observer, but we reasoned that it would not be likely to find observers that would be really “blind” to the logic behind the study. Although the phrase: “we did not inform the observer about the hypothesis of the study” would have sounded great and would have suited the distracted reader, it would have been a hollow one, as such blindness was impossible to obtain. Although our results do not rule out the possibility of some effect of this bias, we took extensive measures to avert the risks of possible negative impacts. First, our dependent variable, littering yes or no, leaves little room for interpretation. More importantly, we established beforehand a clear protocol on what we would consider littering the flyer and what not (as reported in the *Science* publication). Dropping the flyer in the alley, hanging or laying it on another bicycle, putting it in the cycle bag of another bicycle (it happened) was considered littering, everything else was noted as not littering. We made these decisions beforehand, precisely to avoid such biases as rationalizations or flexible interpretations afterwards as to increase the likelihood to find support for our hypothesis. Does this mean that people that were recorded as “not littering” may have littered when they were out of sight, or even after leaving the alley? Yes, this may be the case. Yet, if so, this would be true in both

conditions, making it less likely that it would have seriously affected our results.

A limitation of the study that was not mentioned by the authors is that we did not (and could not) include a manipulation check. Hence, we cannot be sure that the participants in the graffiti, litter, or sign conditions actually saw our manipulations (e.g., the presence of graffiti or litter). We experimented with this before starting the actual study, but it was hard to establish, making it a possible source for the observer bias we tried to avoid. The authors suggest videotaping the scene and later code the data, enabling to establish interrater reliability scores. However, there are two considerable obstacles to videotaping. First, there is practical hurdle: we would have had to have quite a number of cameras from different angles which would have been prohibitive. Second, there are ethical code restrictions of filming people without their consent for purposes of a scientific study.

With regard to measures, the main recommendation of the authors is again that we should have adopted an approach similar to Reiter and Samuel (1980), who simply counted the number of littered flyers after the experiment. This is indeed a very simple method that would leave little room for interpretation, and it is also timesaving. We of course considered it, but discarded it nevertheless, most of all for one specific obvious reason. Our hypothesis states that observing signs of norm-violating behavior such as graffiti or litter makes other norm-violating behavior more likely. Therefore the accumulation of litter that would be the result of this approach would directly influence the likelihood of littering, and be an important confound. Indeed, a study by Cialdini, Reno, and Kallgren (1990) clearly showed that the likelihood of littering goes up when more litter is present in the setting. Given that the presence of litter yes or no was actually one of the manipulations in our study (Keizer et al., 2011, Study 1) we felt that it was very important to keep this constant across participants and conditions. This implied a far more laborious approach: it meant pausing the experiment after every participant that littered, cleaning up the litter, waiting until

everyone had left the setting that might have observed the cleaning action and resuming the experiment.

The two-condition set-up with a simple “yes or no littering” dependent variable made the usage of a chi-square test (or alternatively a z-test for proportions) most appropriate. It also allowed us to express the results in percentages, which (as wanted) greatly contributed to making the work understandable for an audience outside the scientific domain. The chi-square is sensitive for very large samples (i.e., more likely to produce a significant result), we therefore deliberately restricted the number of participants in each study. In Keizer et al. (2011), we added a number of conditions in which we used a relative larger (but still conservative) sample size because we saw in the previous experiments that we possibly needed more power for the prohibition sign effects to show up in this setup (as also explained in the article). This legitimate procedure is quite different from adding subjects to one experiment until the results are significant. Also, the larger sample conditions are presented as separate conditions, apart from the similar smaller conditions to give the reader insight.

Cumulation and Inclusive Replication

We are the first to acknowledge that all our preparations as described before of course do not guarantee the absence of all unwanted effects. In fact, we agree with Burkley and Blanton (2008) and McGrath (1982) that all studies and study designs have strengths and weaknesses, and that it is not possible to design THE perfect study that meets all possible methodological concerns. We therefore tried to address any remaining possible biases and confounds by one key solution (also) proposed by these authors: replication and tests of robustness through cumulation. In doing so, we followed an approach in which we tested the same hypothesis in multiple studies. The clearly formulated hypothesis derived from goal framing specified the direction of the expected cross-norm effect, which allowed one-sided

testing. The decision to use multiple studies to test this hypothesis, helped to avoid ad hoc adaptations of the hypothesis on the basis of previous results. In our line of studies we specifically used different set-ups, with different manipulations and dependent variables, and in different settings. All the experiments in the 2008 paper have their own strengths and weaknesses. The goal of this line of studies was not only to test whether the found effects applied to different norms, but also to increase our confidence that these effects were not the result of the weaknesses of or biases in a single study. In many instances, our studies purposely tried to address shortcomings from the previous studies. In fact, this is an ongoing process that probably never stops. In a set of recent studies in this line, published in *PLOS ONE* (Keizer, Lindenberg, & Steg, 2013) we take on a series of unaddressed questions, some of which are also raised by the authors of the critique. For example, in these studies we explicitly studied whether the effects of signs of norm violations by others are influenced by the presence of other people. For this reason, in one of these studies we solely focused on people that are alone in the chosen setting (which produced similar findings and replicated the effects of the previous studies). We also tested whether groups behave differently than individuals when observing signs of norm-violating behavior; our results revealed that the pattern of results was very similar for individuals and groups. Thus when we explicitly tested for the dependence effect, we did not find it. The 2011 studies had a similar starting point, in these studies we wanted to investigate the effect of the used prohibition sign in our 2008 studies. We specifically wanted to test our hypothesis that the antigraffiti sign strengthened the negative effect of the present graffiti. The mentioned use of the same setting, set-up, and that both studies share this condition is not accidental, but intended to help us answer our question by strengthening the relation between both lines of studies. In a similar respect, the 2013 *PLOS ONE* line of studies starts with a field experiment that is comparable to the last studies of our 2008 work.

Other researchers addressed other remaining questions of our research. For example, Torgler (2010) tested (and found support for) the hypothesized spreading effect of norm violations reported in our 2008 paper by studying various norm violations simultaneously using a large survey dataset (between 16,000 and 35,000 participants from 31 different countries). This correlational study enabled him to control for various individual factors that may affect the likelihood of norm transgressions (e.g., age, gender, formal education, risk aversion, employment status, and religiosity). Moreover, Ramos and Torgler (2012) tested (and found support for) the spreading effect in a field experiment that took place in a work environment. This setting allowed them to control for influences of age and sociodemographic factors such as gender, age, and job characteristics. The study by Reiter and Samuel (1980), which the authors use for their retrospective replication of the potential reversal effect of prohibition signs, revealed some interesting results as well which were not mentioned in the critique. Although they used a conceptually different prohibition sign than we used, as it focuses people on the fine for transgression, results suggest an effect similar to what we found. In their critique, the authors indicate that the Reiter and Samuel study that was run on two different days did not find support for the reversal effect of prohibition signs. This is indeed the case when looking at the cumulative littering rates across the 2 days. However, to the surprise of Reiter and Samuel themselves, litter rates on Day 2 were highest in the condition with the prohibition sign in a littered setting (48% vs. 28% in the litter alone condition), which is much in line with the results we reported. It would be interesting to explore why this effect was not observed on the first day.

Conclusion

We agree with the authors of the critique that our field studies have limitations, mostly with regard to randomization and control. Yet, using standards of lab experiments to judge field experiments misses the point. Much goes on in terms

of careful consideration behind the scenes when one plans a field experiment. The considerations described before are only a small part of the many issues we considered when designing our field experiments. Much is and can be done to limit the disadvantages of field experiments in terms of randomization and control. For those who might be discouraged by the apparent work and time associated with such field studies, we would like to stress that it is also a highly enjoyable process. Designing field studies is a creative challenge that entails a thorough preparation involving numerous decisions, many trials and errors. It involves a continuous comparison of costs and benefits associated with different setups before the actual study is run as to enhance internal validity. In addition, cumulations of findings are the most important criterion. It requires a strategy of *inclusive replications*: strict replications combined with replications in different settings and with different paradigms, both inside and outside the lab, following both experimental and correlational designs, and using different statistical methods to test the hypotheses (cf. Burkley & Blanton, 2008). Inclusive replication is a key strategy to overcome the inevitable weaknesses associated with any scientific study, including field experiments.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

- Burkley, M., & Blanton, H. (2008). The dilemmatics approach to applied research design: A case study of violent video game effects and aggression. In L. Steg, B. Buunk, & T. Rothengatter (Eds.), *Applied social psychology: Understanding and managing social problems* (pp. 87–116). Cambridge, UK: Cambridge University Press.
- Cialdini, R. B. (2009). We have to break up. *Perspectives on Psychological Science*, 4, 5–6. doi:10.1111/j.1745-6924.2009.01091.x
- Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6), 1015–1026. doi:10.1037/0022-3514.58.6.1015
- Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *Science*, 322, 1681–1685. doi:10.1126/science.1161405
- Keizer, K., Lindenberg, S., & Steg, L. (2011). The reversal effects of prohibition signs. *Group Processes & Intergroup Relations*, 14(5), 681–688. doi:10.1177/1368430211398505
- Keizer, K., Lindenberg, S., & Steg, L. (2013). The importance of demonstratively restoring order. *PLoS ONE*, 8(6), e65137. doi:10.1371/journal.pone.0065137
- Lindenberg, S., & Steg, L. (2007). Normative, gain and hedonic goal frames guiding environmental behavior. *Journal of Social Issues*, 63(1), 117–137. doi:10.1111/j.1540-4560.2007.00499.x
- McGrath, J. E. (1982). Dilemmatics: The study of research choices and dilemmas. In J. E. McGrath, J. Martin, & R. A. Kulka (Eds.), *Judgment calls in research* (pp. 69–102). Beverly Hills, CA: Sage.
- Ramos, J., & Torgler, B. (2012). Are academics messy? Testing the broken windows theory with a field experiment in the work environment. *Review of Law & Economics*, 8(3), 563–577. doi:10.1515/1555-5879.1617
- Reiter, S. M., & Samuel, W. (1980). Littering as a function of prior litter and the presence or absence of prohibitive signs. *Journal of Applied Social Psychology*, 10, 45–55. doi:10.1111/j.1559-1816.1980.tb00692.x
- Torgler, B. (2010). Comment on “The Spreading of Disorder.” *Economics Bulletin*, 30(1), 519–523.
- Wicherts, J., & Bakker, M. (in press). Broken windows, mediocre methods, and substandard statistics. *Group Processes & Intergroup Relations*.
- Wilson, J. Q., & Kelling, G. L. (1982, March). Broken windows. *The Atlantic Monthly*, 29–38.