Economics Letters 147 (2016) 51–54

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Policy evaluation, randomized controlled trials, and external validity—A systematic review

Jörg Peters [a,b,*], Jörg Langbein [a], Gareth Roberts [b]

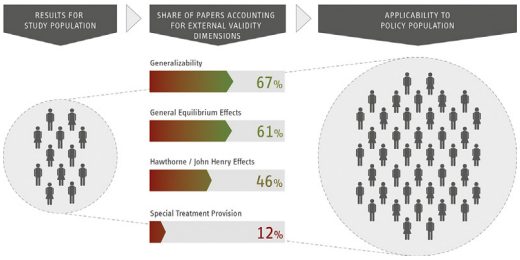
[a] RWI Leibniz Institute for Economic Research, Hohenzollernstraße 1-3, 45128 Essen, Germany

[b] AMERU, University of the Witwatersrand, Johannesburg, South Africa

HIGHLIGHTS

- Systematic review of all RCT-based papers published in top economics journals.
- The review assesses how these articles deal with external validity.
- We find that the majority of papers does not discuss external validity issues.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 17 May 2016
Received in revised form
8 August 2016
Accepted 10 August 2016
Available online 16 August 2016

JEL classification:
C83
C93

Keywords:
Policy evaluation
Internal validity
External validity
Randomized controlled trials

ABSTRACT

This paper reviews all Randomized Controlled Trials (RCTs) published in leading economic journals between 2009 and 2014 with respect to how they deal with potential hazards to external validity: Hawthorne and John-Henry effects, general equilibrium effects, specific sample problems, and special care in treatment provision. We find that the majority of published RCTs does not discuss these hazards and many do not provide the necessary information to assess potential problems.

1. Introduction

External validity prevails if a study's findings can be transferred from the study population to a different policy population. In terms of internal validity, one method stands out: Randomized controlled trials (RCTs). Self-selection into treatment is not a problem due to the randomized assignment of the treatment. The high internal validity of RCTs is frequently contrasted with shortcomings in external validity. Critics state that establishing external validity is in many cases more difficult for RCTs than for studies based on observational data (Dehejia, 2015; Muller, 2015; Pritchett and Sandefur, 2015; Ravallion, 2012).

This paper systematically reviews the extent to which policy evaluations based on RCTs published in top economic journals establish external validity. While there is no uniform definition of external validity in the literature, the seminal toolkit by Duflo et al. (2008) identifies four hazards to external validity: Hawthorne and John Henry Effects, general equilibrium effects, specific sample problems, and special care in the treatment provision. The idea

* Corresponding author at: RWI Leibniz Institute for Economic Research, Hohenzollernstraße 1-3, 45128 Essen, Germany.
E-mail address: peters@rwi-essen.de (J. Peters).

http://dx.doi.org/10.1016/j.econlet.2016.08.013
0165-1765/

underlying our review is that not all RCTs are equally exposed to these respective hazards—which is why a discussion in each paper is needed to establish how externally valid the study's findings are.

2. Why are RCTs more prone to external validity hazards?

It is often argued that high-quality observational studies based on panel data that cover a long period and whole countries or more, achieve a higher degree of external validity than RCTs, which often can only be done in a limited region and have to rely on short period data (see Dehejia, 2015; Ravallion, 2012). Furthermore, the controlled character of RCTs is suspected to co-determine the results in a way that findings cannot be readily transferred to non-study set-ups. More specifically, to the extent that participants in an RCT are aware of their participation in a randomized intervention, they can be expected to behave in a different manner than they would in a scaled and non-monitored intervention. In addition, RCTs are often implemented by small non-governmental organizations (NGOs), which, especially in developing country contexts, might lead to more effective treatments than what can be expected if the intervention is implemented by a governmental agency. Although these concerns do not per se demonstrate that RCTs achieve lower external validity, many RCTs exhibit parameters that justifiably raise more concerns than equally noticed observational studies.

3. Potential hazards to external validity

Following Duflo et al. (2008) we identify four potential hazards to external validity. First, **Hawthorne and John Henry effects** might occur if the participants in an RCT know or notice that they are participating in an experiment.[1] This could lead to an altered behavior in the treatment group (Hawthorne effect) and the control group (John Henry effect). Such behavioral responses clearly differ between different experimental set-ups. If the experiment is embedded into a business-as-usual set up, distortions of participants' behavior are unlikely. In contrast, if the randomized intervention interferes perceivably with the participants' daily life, participants will probably behave differently than they would under non-experimental conditions.[2] Hence, transparency on the experiment's obtrusiveness and a qualitative discussion are needed.

Second, **general equilibrium effects** (GEE) only become noticeable if the program is upscaled to a broader population or extended to a longer term. RCTs that study market outcomes, for example, are more prone to GEE.[3] This is why a profound discussion of potential GEE would be desirable. Third, the **specific sample problem** occurs if the study population is different from the policy population in which the intervention could be brought to scale. A quantitative data-based or at least a qualitative discussion about the particularities of the study population is much desired. Moreover, an examination of effect heterogeneity and a scaling of treatment effects relative to the standard deviation are important to improve the reader's assessment of transferability and the comparability of treatment effects across different studies.

Fourth, **special care** in the provision of the treatment compared to how the intervention would be implemented in a scaled program threatens the transferability of the findings. Bold et al. (2013) and Allcott (2015) provide compelling evidence suggesting the effectiveness observed in RCTs to be higher than what can be expected if the evaluated program is scaled. Therefore, the RCT's treatment provision should be discussed in comparison to how it would be provided in a scaled intervention.

4. The systematic review

We reviewed all RCTs published between 2009 and 2014 in the *American Economic Review, Econometrica*, *Quarterly Journal of Economics*, *Journal of Political Economy*, *Review of Economic Studies, Economic Journal*, *Journal of Public Economics*, and the *American Economic Journal: Applied Economics*. In total, we included all 92 papers in the review that use an RCT to evaluate a policy. Mere test-of-a-theory papers are excluded. Each paper (including the online appendix) was asked ten simple *objective* questions. Our Online Appendix contains a short report on each paper documenting the answers to the questions as well as quotes from the paper to substantiate the answer. As part of our methodology, we sent these reports out to the 73 lead authors of the 92 papers to verify the answers.[4]

Hawthorne and John Henry:

1. Does the paper explicitly say whether participants are aware (or not) of being part of an experiment or a study?

Papers that receive a 'no' for Question 1 do not discuss potential biases resulting from Hawthorne or John Henry effects, because a statement on the participants' awareness of the study is the obvious point of departure for this discussion. Note that unlike lab or medical experiments participants in social science RCTs are often not aware of participating in an experiment.

2. If people are aware of being part of an experiment or a study, does the paper discuss potential implications of Hawthorne or John-Henry effects in the interpretation of the treatment effect and its mechanisms?

General equilibrium effects:

3. Does the paper explicitly discuss what might happen if the program is scaled-up?[5]

4. Does the paper explicitly discuss if and how the treatment effect might change in the long run?

Specific sample problem:

5. Does the paper explicitly discuss the policy population (to which the findings are generalized) or potential restrictions in generalizing results from the study population?

It is sometimes argued that not all RCTs intend to generate generalizable results and are rather designed to test a theoretical concept. We therefore ask a filter question "Does the paper generalize beyond the study population?" and only apply Question 5 to those papers that do generalize. This includes virtually all of the reviewed papers, although to different degrees.[6]

We ask three further questions with respect to the specific sample problem:

6. Does the paper provide a quantitative comparison of the study population and the policy population?

[1] See Bulte et al. (2014) for evidence on strong behavioral response effects including Hawthorne effects in Tanzania.

[2] Cilliers et al. (2015) provide evidence for the distorting effects of foreigner presence in framed field experiments in developing countries.

[3] See Crépon et al. (2013) for an example of such GEE in a randomized labor market program, in which treated participants benefited at the expense of non-treated participants.

[4] Note that Questions 7 and 8 were added after this authors' feedback round. See Peters et al. (2015) for more details on the review and its methodology.

[5] This question does not apply to programs that are already implemented at scale, for example country wide. Only four papers in our review use data based on such a program (all on the Mexican PROGRESA program). We excluded these four papers from Question 5.

[6] In fact, our review focuses on policy evaluations, so indeed 89% of the reviewed papers do generalize, see the Online Appendix for details.