'Misbehaving' RCTs: the confounding problem of human agency

Naila Kabeer

Departments of International Development/Gender Studies,

London School of Economics and Political Science

Email: N.Kabeer@lse.ac.uk

Telephone: +44 791 944 5087

Abstract: *This paper argues that the theoretical model of causal inference underpinning RCTs is frequently undermined by the failure of different actors involved in their implementation to behave in ways required by the model. This is not a problem unique to RCTs, but it poses a greater challenge to them because it undercuts their claims to methodological superiority based on the 'clean identification' of causal effects.*

The 2019 Nobel Prize in Economics was awarded to three leading proponents of randomized control trials (RCTs) on the grounds that their carefully designed field experiments were providing reliable answers to the question of 'what works' in the fight against global poverty. What distinguishes RCTs from non-experimental approaches to this question is the random assignment of subjects drawn from a similar population to treatment and control groups *prior* to an intervention. Its proponents claim that this allows them to avoid the selection biases associated with retrospective evaluations so that difference in the outcomes of interest can be 'cleanly' attributed to the intervention. As others point out, it also minimizes the need for contextual information to identify confounding influences on outcomes or to interpret causal processes (Deaton and Cartwright, 2018; Shaffer, 2011).

The growing dominance of RCTs has not gone unchallenged; critiques encompassing ethical, methodological and political issues. In this contribution, I focus on examples of RCT 'misbehaviour' and their implications for the causal stories that RCTs seek to tell. I borrow the term from an article by Buvinic (1986) on project 'misbehaviour' which explored the reasons why certain projects systematically failed to realize their stated goals. It was, of course, not the projects that misbehaved. Rather, the misbehaviour represented the (frequently unanticipated) agency of stakeholders involved with the project acting on their particular assumptions, interests and constraints or responding to contextual differences in ways that undermined project effectiveness. I draw on studies of three RCTs conducted by leading proponent in the field to discuss the challenge that similar forms of misbehaviour present for their methodological claims.

Two of my examples are drawn from the Graduation Programme co-ordinated by the Innovations for Poverty Action (IPA). This carried out RCTs in six countries to test the generalizability of the Targeting the Ultra Poor (TUP) approach pioneered by BRAC, Bangladesh. The approach combined asset transfers to women in poverty with other supportive measures intended to build their entrepreneurial capacity. A second group of researchers, including myself, were simultaneously commissioned by the funders of

Graduation Programme to carry out qualitative assessments of two other TUP projects being implemented in West Bengal (India) and Sindh (Pakistan) in close proximity to the Graduation RCTs. Despite funders' efforts, IPA refused to consider the possibility of an integrated approach[i]. I was, nevertheless, curious to explore reasons for overlaps and divergences in the findings reported by the two sets of studies (Kabeer, 2019). A close reading of relevant publications brought to light examples of misbehaviour in both RCTs which, I would argue, seriously compromised the usefulness of their findings.

The 'misbehaviour' in the West Bengal RCT was by selected beneficiaries. 36% of the households invited to participate in the 'treatment' group turned down the invitation. They were 'predominantly Muslim', reportedly suspicious of the motives of the implementing agency. The RCT dealt with this by confining their estimates to 'intent to treat' effects of the project. These averaged project's impacts across those who were invited to participate in the treatment, regardless of whether they accepted or not. The researchers justified their focus on the grounds that these estimates '[gave] the expected impact of the project' and '[were] most relevant to the issue of scaling up the programme' (Banerjee et al., 2011: 10).

An alternative estimate would have been 'treatment on the treated' effects which average effects for those who actually participated in the project, arguably providing more realistic measures of impact. The one example provided of both estimates – 15% increase in per capita consumption for all invited households compared to 25% for just those who accepted the invitation – suggested that households that refused to participate reported considerably lower, even negative, impacts. As critics have pointed out, the RCT preoccupation with 'average' effects leads to the neglect of the distributional issues, despite the fact that these have particular relevance to the political economy of 'scaling up' (Bardhan, 2013). In the West Bengal study, the distribution of benefits was systematically skewed against a marginalized religious minority, potentially exacerbating pre-existing social divisions.

In the Sindh RCT, it was those responsible for the randomization process who 'misbehaved'. Five implementing organizations were required to identify the least developed villages within selected locations, select the poorest households through participatory techniques and use a lottery to randomly assign half the selected households in each village to treatment groups and the other half to control groups. Unfortunately, miscommunication meant that while some of the organizations used lotteries, others simply assigned selected households in half of their villages to the treatment group and selected households in the other half to the control group.

This misbehaviour was not discussed in publications by the concerned RCT scholars (Karlan and Parienté, 2014; Banerjee et al. 2015). It was reported instead in an independent evaluation of the Sindh TUP projects by Innovative Development Strategies (2012) commissioned by Pakistan Poverty Alleviation Fund. As it pointed out, the problematic sampling process meant there was no guarantee that relevant characteristics were identically distributed across treatment and control households or indeed that they were equally poor. As it happened, a later publication by IPA/J-PAL (2015) celebrating the 'successful targeting of the ultra-poor households' by the Graduation RCTs reported, without comment, that 82% of the households selected to participate in the Sind RCT started out *above* the poverty line.

My third example draws on an 'insider' account of an RCT on microcredit in Morocco carried out by J-PAL in collaboration with a major microfinance organization (Bédécarrats et

al. 2019). Two of the authors of this account had, at the request of the funders of the project, carried out a qualitative study alongside the RCT. The J-PAL team declined to collaborate. However, the insider status of the qualitative team allowed them detailed insights into the unfolding of the RCT, including examples of misplaced assumptions and unruly practices that undermined the experimental design of the study.

For instance, the plan was to locate the RCT in remote villages assumed to be free of microcredit, thereby allowing isolation of the effects of the 'treatment'. In reality, the selected villages varied considerably in terms of proximity to urban areas and many had access to microfinance organizations. Most of these left halfway through the RCT because of a default crisis in the sector but their presence muddied the impacts captured by the end line survey.

Furthermore, based on experience in urban areas, the RCT anticipated high demand for microfinance for entrepreneurial purpose. In reality, demand was found to be extremely low and heterogeneous, ranging from zero in some villages to 55% in others. This seriously compromised the study's ability to detect small effects when comparing treatment and control households. Consequently, the J-Pal team undertook various 'tweaks' to deal with the problem.

One set of tweaks served to modify the intervention: additional information campaigns were launched to motivate demand; bonuses were introduced to incentivize loan officers; and the minimum quota for women was dropped as men were assumed to have a higher propensity to borrow. The other set of tweaks modified the sampling approach: village borders were altered in the hope of finding more clients, new households were added to the end-line survey and villages with zero take-up were dropped. By the end of the project, it was unclear what or who was being evaluated, since both supply of microfinance and definition of treatment group had undergone considerable change as the experiment progressed. However, the subsequent publication about the RCT (Crépon et al, 2015) failed to refer to any breaches of protocol or to critiques published by the qualitative researchers.

The examples cited here raise a problematic issue. It was only possible to gain insights into 'misbehaviour' within these RCTs by careful re-reading of available materials by an 'outside' researcher and through the 'insider' knowledge of an independent qualitative team. This is in line with the assertion by Barrett and Carter (2010) that, unlike publications in the natural sciences, RCT publications rarely report crucial details about deviations between design and implementation. We therefore do not know how widespread these deviations are or how they affect the validity of reported results. Yet such misbehaviour is common to all field-based evaluations and, as Deaton (2010) notes, there is nothing that grants RCTs special immunity. It is therefore puzzling that RCTs are not more transparent about the problems they encounter and their efforts to deal with them. A plausible explanation, suggested by Bédécarrats et al., is that such transparency might compromise the claims to simplicity and rigour that underlie the 'gold standard' status of RCTs. They would run the danger of becoming just one of many 'good enough' methodologies seeking to reconcile theoretical assumptions with the messiness of the field.

REFERENCES

Banerjee, A.,Duflo, E., Chattopadhyay R., & Shapiro, J. (2011) *Targeting the hard-core poor: an impact assessment* (http://www.povertyactionlab.org/publication/targeting-hard-core-poorimpact-assessment) (Accessed 15-11-17)
Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Pariente, W., Shapiro, J., Thuysbaert, B. & Udry, C (2015) 'A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries', *Science* 348 (6236): 1260799–1260799
Bardhan, P. (2013) Little, big: two ideas about fighting global poverty *Boston Review* May 20th. http://bostonreview.net/world-books-ideas/pranab-bardhan-little-big (Accessed 20-11-19).
Barrett, C.B. and Carter, M.R. (2010) The power and pitfalls of experiments in development economics: some non-random reflections *Applied Economic Perspectives and Policy 32* (4): 515–548.
Bauchet, J., Morduch, J. and Shamika, R. (2015). Failure vs. displacement: why an innovative anti-poverty program showed no net impact in South India. *Journal of Development Economics* 116 (2015): 1-16
Bédécarrats, F., Guérin, I., Morvant-Roux, S. and Roubaud, F. (2019) *Lies, damn lies and RCT: A J-PAL RCT on microcredit in rural Morocco* Working Paper DT 2019-04. Paris: UMR Dial 225
Buvinic, M. (1986) 'Projects for women in the Third World: explain their misbehaviour' *World Development* 14 (5): 653-664
Crépon, B. Devoto, F. Duflo, E and Parienté, W. (2015) Estimating the Impact of Microcredit on Those Who Take It Up:Evidence from a Randomized Experiment in Morocco. *American Economic Journal: Applied Economics* 7(1): 123–150
Deaton, A. (2010). 'Instruments, randomization, and learning about development' *Journal of Economic Literature* 48: 224-255
Deaton, A.& Cartwright, N. (2018) Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine* 210 (2018): 2-21
Innovative Development Strategies (2012) *Assessment Survey: PPAF's Social Safety Net Targeting Ultra Poor (TUP) Program* (http://www.ppaf.org.pk/Research/TUP.pdf). Accessed 3-9-17)
IPA/J-Pal (2015) *Building stable livelihoods for the ultra-poor* Policy Bulletin September (https://www.povertyactionlab.org/policy-insight/building-stable-livelihoods-ultra-poor. (Accessed 14-10-18).
Kabeer, N. (2019) Randomized control trials and qualitative evaluations of a multifaceted program for women in extreme poverty: empirical findings and methodological reflections. *Journal of Human Development and Capabilities* 20 (2): 197-217
Karlan, D. and Parienté, W. (2014) *Impact of 'Targeting the Ultra-Poor' Program in Pakistan* Research Partnerships on Participatory Development Brief 9. PPAF: Islamabad. (http://www.ppaf.org.pk/PPAF-Conference-2014/Publications/paper9.pdf). Accessed 3-5-15.
Shaffer, P. (2011) 'Against excessive rhetoric in impact assessment: overstating the case for randomized controlled experiments' *Journal of Development Studies* Vol. 47 (11): 1619-1635
White, H. (2009) 'Theory-based impact evaluation: principles and practice' *Journal of Development Effectiveness* Vol. 1 (3): 271-284

[i] Although the Graduation Programme commissioned quantitative and qualitative studies, the only RCT of a TUP pilot that attempted to integrate the findings of a qualitative study was carried out independently of the IPA (Bauchet et al, 2015). Some RCT researchers have long advocated the need to combine RCTs with other methods (White, 2009), but progress remains extremely slow.