The European Journal of Development Research (2019) 31:169–173
https://doi.org/10.1057/s41287-019-00199-2

COMMENTARY

A Commentary to 'Bridging to Action Requires Mixed Methods, Not Only Randomised Control Trials'

Marie Gaarder[1]

Published online: 25 March 2019

Let's start with a confession: I agree with many points made in this article. The problem—not uncommon in this kind of positioning pieces—is that it creates a caricature version of a method in order to then prove that its arguments and methods are better. Most methods can be carried out poorly or thoughtfully, and RCTs are no exception. What bothers me is that RCTs tend to be picked on. There seem to be two reasons for this: they are more clearly understood and defined than arguably any other evaluation method, and they have experienced a boom in funding in recent years, which apparently invites envy. I would argue that these same reasons for picking on RCTs may be some of their biggest contributions to the evaluation field! Experimental and quasi-experimental methods set out clear minimum quality criteria, reporting criteria, and criteria to assess the risk of bias AND can be replicated by others and hence do not rely on non-transparent 'expert judgements'.[1] This clarity, transparency and replicability is one of their substantial contributions to the evaluation field and an example that other methods should strive to follow. We in the International Initiative for Impact Evaluation (3ie) have recently launched our Research Transparency Policy, which supports independent replication of reported quantitative results and the transparent sharing of de-identified data and codes. Given our focus on theory-based, mixed-methods impact evaluations, the next challenge is to similarly enhance the transparency and public availability of the qualitative data and findings upon which evaluations draw. We hope that other organizations and researchers will join us in this endeavor.

Now to the second reason for picking on RCTs. The author of the lead piece feels that funding has been disproportionately favoring RCTs, and impact evaluations more broadly. She seems to assume a zero-sum, static game, whereas I would venture that the credibility of the evaluation field has been enhanced, which

[1] Olsen (2019) argues that this transparency which includes registering protocols and pre-analysis plans prevents adaptive research, but this is not the intention nor the case—rather, researchers can update their plans with a good justification but at least there is a logged and transparent record that prevents adapting the research question to the data.

✉ Marie Gaarder
mgaarder@3ieimpact.org

[1] International Initiative for Impact Evaluation (3ie), New Delhi, India

has increased demand and funding. The arrival of more rigorous methods to discern attribution and this broader pallet of tools and methods, used appropriately by skilled evaluators, enhances the utility of evaluations. This should increase demand for evaluations in general, and therefore funding. Or to use a counterfactual, if the funding for evaluations does not increase—and there are some worrying signs that this will be the case—it is likely that it would have decreased more rapidly had the evaluation field not constantly tried to get better at what it is doing.

Now, having got that off my chest, the fact the article makes abundantly clear and with which I wholeheartedly agree, is that no single method should be used to identify and respond to complex evaluation questions; rather, a whole range of methods are available to answer different evaluation questions and the best combination of methods will depend on the questions prioritized and the data at hand.

I will make two points in my commentary: first, I will argue that the impact evaluation methods supported by the International Initiative for Impact Evaluation, 3ie, and increasingly other organizations (we happily claim contribution to this movement!) are entering the era of 'evaluation for grown-ups' and are tackling many of the issues that this article raises. Thereafter, I will bring up some additional challenges that we are tackling and that the article has not discussed.

Since its first evidence program was launched 10 years ago, 3ie has insisted that all the impact evaluations it funds and supports should draw upon mixed methods, should fully spell out the underlying theory of change to be tested and revisited, should involve and engage key stakeholders throughout, and should perform cost-effectiveness or cost–benefit assessments. A recent paper by Jimenez et al. (2018) reviews a sample of IEs from numerous IE repositories to explore the ways in which methods are being mixed and to identify good practices of integrating qualitative methods into quantitative impact evaluations (IEs). They find that the IE studies that used mixed methods mainly did so at the tail-end of the study in order to help interpret the effect findings.[2] Furthermore, while methods to account for bias were generally well described for quantitative components of the impact evaluations, fewer studies were found to demonstrate comparable thoroughness with the qualitative components. For instance, only 20% of studies reported on the analytical framework for qualitative data; only 38% of the studies presented information on their qualitative sampling; and only 20% of the studies reported any form of validity checks for their qualitative findings. In summary, studies mostly do better on quantitative rather than qualitative rigor. They find that successful mixed-methods impact evaluations provide a clear rationale for the integration of methods, deploy multidisciplinary teams, provide adequate documentation, and acknowledge the limitations to the generalizability of qualitative and quantitative findings. Successful integration tended to improve the evaluations by strengthening data collection and validation, analysis,

[2] An interesting question that could be further explored is whether the mixing of methods at the tail-end was equally distributed between expected and unexpected findings. In my experience, we more rarely see researchers ask why something worked; there seems to be a clear bias towards explaining unwanted or unexpected findings whereas it could equally be that we get the desired or expected findings for reasons others than those we assume.

interpretation and policy recommendations. To summarize, mixed-methods impact evaluations are increasingly the standard, but the mixing needs to be more fully built into all phases of evaluation implementation, the teams need to have the right skills mix, and the 'other' methods need to have clearer quality and reporting guidelines.

Another aspect shows that we have entered into the era of evaluation for grown-ups: there are very few voices left out there that still have the attitude of 'I have a hammer hence everything I see are nails' (or 'I know how to do RCTs and am looking for something to randomize'). Increasingly, 3ie and others are adopting a more holistic approach to evaluation. This holistic approach can be summarized by the following set of principles:

- Evaluations are time and resource intensive and hence should be carried out only to fill important evidence gaps.
- Implementing rigorous causal evaluations (impact evaluations) to establish relative or absolute effectiveness of development interventions only make sense after having established: (1) evaluability; (2) take-up and relevance of the intervention; and (3) implementation fidelity. In 3ie, what this means is that our evidence programs often build in a formative and process evaluation phase in order to assess these factors and identify the impact evaluations with the most value for money.
- We aspire to use the best available data to answer the questions most relevant to policy makers, rather than the perfect data to answer irrelevant questions. This also entails faster learning loops, research around implementation issues, etc.
- In order for evaluations to contribute to improvements on the ground, a range of efforts need to take place, including: (1) involvement of local researchers in the evaluation efforts; (2) extensive stakeholder engagement throughout; (3) efforts to measure similar indicators across studies and to look at comparable interventions across contexts for comparability purposes; and (4) addressing the limited generalizability of individual impact evaluations by producing systematic reviews which appraise and synthesize all the available high-quality evidence on the effectiveness of specific social and economic development interventions in low- and middle-income countries.

So far, I have argued that recent developments in the field of RCTs and impact evaluations more broadly show that many aspects of the criticism in the Olsen-paper are already in the process of being addressed, though much remains to be done to improve the quality of integration of methods.

My second contribution is around limitations to the implications that can be drawn from impact evaluations. When not sufficiently recognized, these can lead to misleading conclusions and add fuel to the fire for the IE sceptics. The limitations are three-fold: misdiagnoses, confounding implementation issues and design issues, and imperfect understanding about the effectiveness production function. I will treat each in turn.

Development interventions have similarities to medical treatments: if you treat superficial symptoms rather than the underlying pathology, or if you give the wrong medicine, you will not cure the illness. In medicine, you would not say that the medicine was ineffective in general, you would say that the doctor

misdiagnosed the pathology. Similarly, in international development, we can only judge the effectiveness of an intervention after we ascertained that it was designed to address the main underlying problem or 'binding constraint'. Yet, all too often, as impact evaluators we judge the effectiveness of development interventions without knowledge of whether policy makers and aid agencies established the correct diagnosis of the root cause (or causes) of a certain development problem prior to designing an intervention to address it. As I have argued elsewhere, misdiagnosis is a widespread phenomenon in international development. An example: conditional cash transfer (CCT) programs have successfully increased the frequency with which poor people had health check-ups and improved school attendance of poor families' children. Alas, the evidence that CCTs have improved education and health outcomes has been mixed at best (Gaarder et al. 2010; Snilstveit et al. 2015). The reason could be misdiagnosis of the bottlenecks for improving human capital outcomes. These may have been on the supply side of the services (no medicine, no teachers, etc.). Therefore, there was a mismatch between diagnosis and treatment, and we should not conclude that CCTs are ineffective at improving health and education outcomes in general. Indeed, if the health centers and schools had been adequately staffed, resourced and trained to service the additional demand, and if non-attendance had been verifiably due to the costs, then CCT programs would likely have led to improved health and education outcomes.

The right diagnosis is a necessary condition to achieve impacts in development interventions. Nevertheless, scant attention is paid to this in practical impact evaluation work and hence the wrong conclusions are often drawn. An impact evaluation researcher who does not have sufficient knowledge about the diagnostic work that went into designing an intervention or program, also does not know whether the lack of effect found was due to the intervention being ineffective (in general) or because the development practitioners misdiagnosed the problem. In recent promising changes to systematic review methods (in '*Do participation and accountability improve development outcomes? A systematic Review*' by Hugh Waddington, Ada Sonnenfeld, Jennifer Stevenson and team, forthcoming), 3ie is categorizing impact evaluation studies according to the assessment of underlying developmental bottlenecks.

As for the need to disentangle implementation issues from design issues, this is an obvious point that nevertheless frequently crops up in evaluation reports. An example from an impact evaluation carried out on the Honduras' CCT program, PRAF, is that the multi-arm clustered RCT found that the supply-side interventions had no effect on the desired human capital outcomes. They concluded that the supply-side interventions had indeed not been implemented due to some procurement obstacles. What is wrong with this picture? It would have been much more cost-effective to first check whether the program was being implemented as planned (implementation fidelity) through a process evaluation, and, once it was confirmed that it was not, either the IE should have been postponed until such time as it made sense or the treatment arm that related to the supply-side should have been dropped from the IE and substantial amounts of resources could have been saved.

As for the understanding or lack thereof of the effectiveness production function, this one is equally worrying. To date, most effectiveness studies (impact evaluations) are one-off in the sense that they measure the attributable change of some desired outcome indicator over one period of time, the time period usually set so that we should expect to see some movement. However, impacts of the intervention may change over time and are likely to be non-linear, and hence findings will be very sensitive to the point in time at which impact is measured. For example, for projects that try to increase the participation and empowerment of marginalized groups the literature suggests that the most likely shape of such projects' impact over time is a J curve; that is, things get worse before they get better.

As a conclusion to this second part of my contribution, the impact evaluation field needs to strive toward a much better understanding of design issues, implementation issues, and the trajectory of impacts over time, and until these are all well understood should be exceedingly cautious in giving bold policy recommendations.

References

Gaarder, M., A. Glassman, and J. Todd. 2010. Conditional cash transfers and health: unpacking the causal chain. *Journal of Development Effectiveness* 2 (1): 6–50.

IFPRI. 2010. Conditional cash transfers in latin America. eds. Michelle Adato and John Hoddinott. Baltimore: The Johns Hopkins University Press.

Jimenez, E., H. Waddington, N. Goel, A. Prost, A. Pullin, H. White, S. Lahiri, and A. Narain. 2018. Mixing and matching: using qualitative methods to improve quantitative impact evaluations (IEs) and systematic reviews (SRs) of development outcomes. *Journal of Development Effectiveness* 10 (4): 400–421.

Olsen, W. 2019. Bridging to action requires mixed methods, not only randomised control trials. *The European Journal of Development Research*. https://doi.org/10.1057/s41287-019-00201-x.

Snilstveit, B., J. Stevenson, D. Phillips, M. Vojtkova, E. Gallagher, T. 'Schmidt, H. Jobse, M. Geelen, M. Pastorello, and J. Eyers. 2015. *Interventions for improving learning outcomes and access to education in low− and middle−income countries: a systematic review*, 3ie Systematic Review 24. London: International Initiative for Impact Evaluation (3ie).

Marie Gaarder is the Director of Evaluation at the International Initiative for Impact Evaluation, 3ie. Previous positions include as Manager in the World Bank's Independent Evaluation Group (IEG), Director of the Evaluation Department of the Norwegian Agency for Development Cooperation (NORAD), and Senior Social Development Economist at the Inter-American Development Bank. Marie has over 18 years' experience managing development evaluation and research programs and operations. Her publications cover a range of topics including cash transfer programs, evaluation in fragile and conflict-affected states, the institutionalization of government evaluation, and the use of evidence in decision-making. Marie holds a PhD in Economics from University College London.