Review for the fteval preprint

Much promise, little practice: counterfactual impact analysis in Austrian RTI evaluation

Authors: Jakob Kofler, Jürgen Streicher, Verena Régent, Isabella Wagner

Reviewer: Peter Kolarz

This paper makes an important contribution by highlighting the low prevalence of counterfactual and causal methods in R&I evaluations. This low prevalence (compared with other policy fields, as the paper notes) is well known, but it is helpful to have a clear view on this specifically in the Austrian case. Further, the paper also makes an important contribution by highlighting the various tensions from different stakeholders leading to competing interests in terms of what kinds of methods and findings are desirable from R&I evaluations. Both these sets of insights are well substantiated and they reflect the wider literature, as well as my own professional experience.

I cannot comment in detail on the methodological approach around case identification. My only question here is what we can make of the five additional studies flagged by stakeholders that were not in the fteval repository (any idea what the 'true' figure of non-fteval-listed evaluations might be, and whether there might be any type of bias in the fteval+5 sample? Are they likely to be more robust than non-listed ones? Less?).

However, I am concerned that a little too much is made in terms of the more detailed conclusions: at just 17 causal studies from a repository of over 700, we are dealing with very low numbers, so I am not sure (1) whether the direct comparison with Germany holds much water, or (2) whether we can really say that the ability to conduct causal approaches is genuinely increasing based on just two recent studies that do so. The figures certainly show that counterfactual/causal methods are extremely rare and that Austria seems to be more or less in line with Germany, but any more specific observations I think should be tempered or qualified, given the overall low numbers, and the possible bias issue in the sample (see above).

This brings me to a second main point I'd like to note about this article: it states openly that it wishes to be 'descriptive and diagnostic rather than advocatory'. I think this is a missed opportunity, as this paper does very well (see above) at highlighting competing pressures around evaluation methods but then doesn't really set out recommendations that adequately seek to resolve them. I'm unsure how prescribing greater competence among evaluators and commissioners resolves the tension between some actors' desire for more realist methods to enable learning and other actors' need for hard quant findings to help demonstrate value of the evaluated programmes. Given the finding that well-executed

causal methods are very rare (and for good reasons), it seems like there is a teachable moment here, and good cause for a more critical approach that makes unrealistic expectations more explicit: does the system's obsession with counterfactual analyses maybe undermine evaluators' ability to offer more formative/realist evaluations capable of helping us all develop better, more successful programmes?

So in short, I think this is a valuable contribution, though some of the detailed observations should be tempered. The discussion could do with being far broader and more critical, as this piece really begins to expose some important tensions around evaluation evidence: who needs what, why, and how realistic and/or damaging are those demands? This paper should certainly be used to kick off wider discussions about these questions!