Econ Polit (2016) 33:83–98
DOI 10.1007/s40888-016-0027-1

RCTs for better policy? The case of public systems in developing countries

Sunil Mitra Kumar[1]

Received: 10 August 2015 / Accepted: 3 March 2016 / Published online: 19 March 2016

Abstract This paper considers the application of randomised controlled trials (RCTs) to improve public systems in developing countries. Arguing that existing critiques of RCTs as to problems with extrapolation and narrowness of scope are especially relevant in this context, I consider the claim that these shortcomings can be ameliorated through better causal explanations. I analyse how theoretical mathematical models are used to construct causal explanations, and argue that it is still difficult to extrapolate or address the subjectivity inherent in the choice of interventions. I illustrate these arguments using two prominent RCTs that have trialled interventions to improve government schools in India.

Keywords Randomised controlled trial · Causal inference · Theoretical model · Public system · Development economics

JEL Classification B41 · C90 · O20

1 Introduction

Randomized controlled trials (RCTs) are advocated as an ideal tool for evaluating social policies (Haynes et al. 2012) because they can enable unbiased estimation of treatment effects. This advocacy extends to contemporary development economics (e.g. Banerjee and Duflo 2011), where the design of effective policy often aims to improve public systems.

This paper takes forward existing arguments on the nature of RCTs to develop a critique of their use as the preeminent research tool for improving public systems in

✉ Sunil Mitra Kumar
sunil.kumar@kcl.ac.uk

[1] India Institute, King's College London, Strand, London WC2R 2LS, UK

developing countries. Debates on the usefulness of RCTs centre on concerns about internal and external validity (Worrall 2007; Cartwright 2007, 2011). There is also the tendency to ask narrow questions (Reddy 2012; Rodrik 2009), leading to policy prescriptions that are circumscribed in ways seldom articulated (Favereau 2014).

Our focus are the twin forms of extrapolation required to get from experimental results to effective policy: from experimental to target context, and from the intervention tested to the version that actually manifests in policy. The former is concerned with the problem of external validity, while the latter addresses the problem of asking narrow questions. The first part of our argument demonstrates why both types of extrapolation are particularly tenuous in developing country contexts when the aim is to improve public systems. Test contexts usually bear structural differences to a target public system, because the latter involve governance structures and multiple layers of actors that test contexts do not. Furthermore, the feasibility requirements for RCTs lead to important differences between tested interventions and emergent or recommended policies.

Both arguments, however, are not absolute. Understanding the causal processes underlying responses to a tested intervention could help extrapolate to a different but related policy, and to a structurally distinct context. As Deaton (2010a, 2010) has argued, the belief that results in an experimental setting will hold across different contexts could be strengthened by explaining the associated causal mechanism, and how it functions in the settings to which extrapolation is desired. Similarly, by making explicit any underlying assumptions, causal explanations can aid informed guesses about a larger set of interventions similar to the one actually tested. Indeed, in the absence of theory, the 'radical skepticism' that threatens the internal validity of observational studies also threatens the external validity of experimental results (Stokes 2014; Barrett and Carter 2014). Therefore, the second part of my argument examines this claim: whether causal explanations in the form of models can strengthen extrapolation from test to target context, and from tested intervention to actual policy.

I illustrate these arguments using two prominent papers in the RCT literature, viz. *Remedying Education: Evidence from Two Randomized Experiments in India* by Banerjee, Cole, Duflo and Linden (2007; hereon BCDL) and *Incentives Work: Getting Teachers to Come to School* by Duflo, Hanna and Ryan (2012; hereon DHR). Both studies report on RCTs to improve government schools in India, and offer corresponding policy prescriptions. They illustrate contemporary practice in development economics where the goal is to evaluate policies that can improve a public system. And they demonstrate the arguments by which an intervention is chosen, how results are analysed and conclusions inferred, and how these are used to make policy recommendations. While BCDL is a purely empirical study, DHR provide a theoretical model as causal explanation. This model motivates both the choice of intervention and the interpretation of results, and enables us to observe the ability of models to address extrapolation challenges.

The remainder of the paper is structured as follows. Section 2 summarises the BCDL and DHR case studies and their appropriateness to illustrate my arguments. Section 3 reviews existing perspectives on the kinds of research questions amenable to RCTs, and explains how these apply to public systems in developing

countries. Section 4 focuses on the challenges of extrapolation in the context of such public systems, examining DHR and BCDL's findings, inference, and emergent policy recommendations. Section 5 then examines the case for theoretical explanations as a way of mitigating the criticisms surrounding extrapolation by delving into the nature of theory and theoretical models in economics. Section 6 concludes.

2 Case studies

Public systems of education in developing countries are frequently the subject of debates about quality and efficiency. Some of the important challenges these systems face include poor student learning outcomes (UNESCO 2005), poor nutrition and mortality among children (Black et al. 2003), and high teacher absenteeism (Kremer et al. 2005). A growing literature has analyzed the effects of reform-oriented interventions tested through randomised trials in these contexts. One category of interventions involves providing materials and other inputs. For instance, Miguel and Kremer (2004) measure the impact of deworming medication on school participation, and Glewwe et al. (2004) examine whether flipcharts improve test scores. A second category of intervention takes the form of monetary incentives for teachers (Muralidharan and Sundararaman 2011) or for students (Blimpo 2014).

Within this literature, I believe that BCDL and DHR's approach and analysis are sufficiently characteristic so as to be useful as case studies, and the positive attention they have garnered in academic and development fora highlights their influence.[1] Both studies focus on the India's government schools, test distinct interventions in different settings, yet offer very similar policy recommendations. BCDL test an input-based intervention, introducing an auxiliary teacher into the classroom, while DHR test an incentive-payment scheme aimed at cutting teacher absenteeism. Their policy recommendations are system-wide, in that they recommend the hiring of contractual teachers, with implications for teacher recruitment and remuneration policies, and the role of teacher training institutions and qualifications frameworks. This systemic nature of their recommendations enables us to examine the nature of inference from results to policy recommendations, and the type of extrapolation required to believe in their external validity.

2.1 The Balsakhi study

BCDL report the effects of two classroom-based interventions on student achievement scores. The main intervention focused on remedial teaching, wherein an auxiliary teacher—the Balsakhi (literally 'child's friend')—worked with children who were performing worse than their peers on basic skills, teaching them as a

[1] BCDL have received wide publicity and evinced interest as a case study in the UNESCO's Education For All report (UNESCO 2005) and the World Bank's World Development Report (World Bank 2003). Likewise, DHR have been hailed for displaying good practice (Guerrero et al. 2012; Armstrong 2006), and for their provision of a theoretical model that explains teachers' behavior (Deaton 2010a, p. 449).

separate group during regular school hours.[2] The RCT was conducted in municipal schools in two cities in western India. Unlike regular teachers, Balsakhis were young women from the local community typically educated up to secondary school. They were given two weeks of training and ongoing support by a Non-Government Organisation (NGO) that ran the programme.

Schools were randomly allocated to treatment, and in treatment schools the intervention took place in grade 3 or 4, while in control schools teaching proceeded as usual. After the first two years of the programme, students in treatment schools performed better than their control-school peers, almost entirely due to an increase in test scores for the students taught by the Balsakhis. Only a small part of the increase in test scores was found to persist a year after the intervention had ceased, a finding that the authors discuss in detail, albeit without a conclusive recommendation. I discuss these findings and the resultant policy recommendations in greater detail below, in Sect. 4.1.

2.2 The camera study

DHR report the effects of an incentive-payment intervention aimed at cutting teacher absenteeism. The study was conducted in rural Non Formal Education (NFE) centers run by an NGO in the state of Rajasthan. These are single-teacher centers for children who do not attend regular school, and the NGO aims to have these children enroll in a government school upon completion of the NFE programme. Under the intervention, teachers' salaries were linked to their absenteeism. In addition to a fixed sum, they received a per-day payment if they were present in class for more than ten days in the month. NFE centers were randomly allocated to treatment, and control group teachers received a fixed sum each month irrespective of attendance. To establish attendance, teachers used a camera to take date-stamped photographs of themselves together with all the children present, at the beginning and end of each day.

To interpret any change in absenteeism behavior, DHR propose a principal-agent model as a causal explanation. In this model, teachers are imperfectly monitored by the NGO, and both parties seek to maximise their respective utilities. Each day, teachers decide on whether to work by trading off the utility gained from leisure if they skip work against the loss in payment they would incur, besides a small risk of being fired or reprimanded. The NGO gains utility from teachers being present because it values children's learning, and trades this off against the salaries it must pay to teachers. This model is discussed in greater detail below (Sect. 5.2) where I relate it to the nature of economic theory more generally.

[2] This (the Balsakhi) intervention forms the basis for the authors' policy recommendations. The second intervention was a computer-aided learning programme which also led to higher test scores for students, but was substantially more expensive than the Balsakhi.

2.3 Locating the interventions

The policy focus for both studies is the Indian government school system. It employs 4.6 million teachers in over a million schools (NUEPA 2014), and has widely acknowledged shortcomings in the quality of teaching and learning provided. For example, as of 2005, 53 % of children had dropped out before completing grade VIII (NCERT 2006), and as of 2012, nearly one in every five children enrolled in grade one failed to complete primary school (NUEPA 2014).

An extensive system of academic and executive infrastructure supports the provision of school education. The academic part consists of the National and State Councils for Educational Research and Training (NCERT and SCERTs, respectively) and District Institutes for Education and Training (DIETs). Together, these institutions are responsible for developing curricula and teaching-learning materials, teacher training, academic support and research. The executive infrastructure consists of directorates and secretariats of education in each provincial state. These multiple layers and inter-state differences result in variation in the quality of government schooling (De et al. 2011). Overall, education is the joint responsibility of state and federal governments, with the majority of funding provided by state governments together with periodic federally-funded schemes.

This level of complexity is, I would argue, characteristic of similar public systems such as health and various types of governance. Therefore, it is a useful site to interrogate the ability of RCT interventions to improve, or at least take into account systemic structures if these interventions are to yield sustainable improvements once translated into policy.

3 Questions for methods

RCTs, and experiments more generally, seek to analyze the effects of causes (Holland 1986). Using them to address problems in developing countries, researchers first describe a particular problem, and then suggest a potential solution in the form of an intervention that needs testing.[3] This approach is useful provided the problems themselves are understood well enough to allow hypothesizing ways of addressing them. Even with such understanding, RCTs are subject to implementation constraints that might limit the sorts of interventions actually tested (Reddy 2012).

Public systems are a case in point, where a single such system might rely on several interlinked factors. These can include the availability of financial resources, physical infrastructure such as buildings, transport and electricity, an atmosphere of work ethic and a lack of corruption. Any assessment of the whole system must take into account all constituent attributes, else policy changes might not yield expected

[3] This paper does not focus on the behavioural underpinnings of RCTs in social contexts, but it is worth noting that RCTs here are often behaviourally-motivated. The wider debate here includes the question of 'nudge' vs. 'boost' understandings of human decision making (e.g. Grüne-Yanoff and Hertwig 2015). Indeed, Davis (2013) argues that several RCT interventions in behavioural development economics are 'nudges', and as paternalistic policy prescriptions are liable to social and cultural imperialism.

benefits in the presence of, say, poor procurement policies that are prone to corruption. An RCT aimed at improving a public system would typically focus on changing a single attribute (e.g. teacher absenteeism) within this system, and gauge how this perturbation changes some set of outcomes (e.g. learning outcomes). This approach tends to neglect those parts of the system that influence the outcome less directly, and the interlinkages between these parts (e.g. teacher training institutions, teachers' qualifications, school curricula and examinations). Time and feasibility constraints also restrict the choice of interventions. For instance, changes to teacher education curricula or recruitment policy take long to manifest, and are difficult to implement experimentally because they require sustained political consent and cross-departmental collaboration. Consequently, as a recent review by Kremer et al. (2013) demonstrates, they are unlikely to be chosen for RCTs.

Justifying their choice of interventions, BCDL explain that children do not learn well in government schools because (a) they are first generation learners lacking parental support; (b) curriculum and pedagogy are unsuited to their needs; (c) 'the school system continues to operate as if it were catering to the elite'. Having discussed mixed evidence on the usefulness of inputs such as textbooks, they suggest that '...inputs specifically targeted to helping weaker students learn may be effective' before introducing the Balsakhi.

Similarly, DHR describe the problem of teacher absenteeism in government schools, which is hard to tackle because 'teachers are a powerful political force, able to resist attempts to enforce stricter attendance rules' due to which 'many governments have shifted to instead hiring "para-teachers"' (p. 1241). Presenting a theoretical model that parametrizes teachers' response to financial incentives, they suggest a payment-incentive scheme as an intervention to curtail absenteeism.

The 'para' teachers referred to are contractually-appointed teachers who are paid a fraction of the salary that regular teachers receive (Govinda and Josephine 2004). While the latter are typically graduates with a teacher training qualification, para teachers are school graduates with minimal teacher training. Owing to fiscal pressures, certain Indian provincial state governments have recruited para teachers from the 1990s onwards (Kingdon and Sipahimalani-Rao 2010), and they constitute nearly a tenth of all government teachers (NUEPA 2014).

This concurrent recruitment of para teachers highlights a tension in policy. On the one hand, policy guidelines—e.g. the National Policy on Education 1986 (Government ofIndia 1998) and National Curriculum Framework 2005 (NCERT 2006)—ask that teachers be viewed as professionals, with rigorous selection, certification and remuneration. On the other, a low-skill, low-pay view of teachers lends credence to an instrumental, cost-minimizing view of education (Kumar et al. 2001) which has been criticised for suggesting that 'anyone can teach' (Halperin and Ratteree 2003).

In other words, the debate on how best to improve government schools has an important ideological divide, and viewing it agnostically, neither approach can be ruled out a priori, and both ought to be empirically evaluated. Yet only one of these approaches is feasible for randomisation, since testing whether teachers-as-professionals work better requires systemic changes that cannot be implemented locally. Thus, we contend that BCDL's choice of para teacher-like Balsakhis and

DHR's advocacy for para teachers represents an important normative judgement which is brought about at least in part by their choice of method. Similar arguments would apply to public systems of health and governance, with the more general contention that preferring RCTs biases enquiries into, and efforts to reform, public systems. In the case of developing countries, fundamental questions about how to organise public systems are often still open to debate—as the para-teacher case illustrates—when these systems are still in the process of being built. This makes any bias arising from the choice of research method even more significant.

I now turn to the question of external validity from experimental results. I first examine existing arguments to explain why concerns about the twin forms of extrapolation referred to above—test to target, and tested intervention to actual policy—are particularly relevant to public systems.

4 From results to policy

Worrall (2007) and Cartwright (2007, 2011) have explained why findings from RCTs should not be ranked any higher than other forms of evidence. Briefly, this is because the process of moving from empirical evidence to a useful causal conclusion requires the test population to be suitably similar to the target population for which the conclusion is desired. Even with perfect internal validity—perfect randomization, and the absence of Hawthorne and John Henry effects—external validity does not follow deductively. It remains dependent on the untestable assumption that the same causal process will unfold in the target population as did in the test one.

These arguments are particularly pertinent when the goal is to use RCTs to inform policy for public systems. As in the case of BCDL, the test population could be a subset of the target system. Alternatively, as with DHR who test their intervention on NFE centres run by an NGO, the test context might be altogether outside the target system. In both cases, the differences between test and target are *structural*, and not only those of two distinct sets of subjects. The target population is a superset of the test one, consisting of institutions, administrative structures and their interlinkages, that is qualitatively distinct from the test site. Therefore, extrapolating to the target population involves not only assumptions about why the same causal process will work elsewhere (e.g. teachers in a different province to the one where the RCT took place), but that this will work even after other parts of the system become involved (e.g. institutions and bureaucrats hitherto uninvolved).

We now examine these steps for DHR and BCDL: the process of inferring from empirical results, providing a causal explanation, and using this explanation as the basis for policy recommendations.

4.1 The Balsakhi intervention

BCDL find that test scores went up in treatment schools relative to control schools, and attribute this to the remedial teaching by Balsakhis. They conclude that '[I]n terms of cost for a given improvement in test scores, scaling up the Balsakhi

programme would thus be much more cost effective than hiring new teachers (since reducing class size appears to have little or no impact on test scores)' (p. 1263). And, that '...these results suggest that it may be possible to dramatically increase the quality of education in urban India...' (p. 1263). That said, at no point were additional regular teachers recruited, nor were they compared directly with Balsakhis (e.g. through remedial teaching by regular teachers). Instead, the authors found that test scores did not improve for the 'not lagging behind' children who did not work with the Balsakhis in treatment schools. These children were in effect 'treated' (authors' quotes) with smaller class size, since their lagging-behind peers were working separately with the Balsakhi. Since this 'treatment' did not improve test scores, the authors infer that hiring additional teachers would not be useful.

Accepting these policy recommendations requires different kinds of assumptions. The first relate to the choice of intervention itself: specifically, the comparison with its closest alternative: simply hiring more government school teachers. BCDL's logic for preferring the Balsakhi has two parts. First, that hiring more government teachers would lead only to smaller class sizes. Since they found that smaller class sizes had no effect on 'non lagging behind' children's scores, they infer that the same would likely hold true for smaller class sizes in general (where both lagging and not-lagging children are taught together). And second, they implicitly assume that remedial teaching by regular teachers would be less cost effective compared to Balsakhis.

Notwithstanding, extrapolation from test to target context and intervention to actual policy based on BCDL's results requires an important assumption: that hiring more Balsakhis will leave the behavior of existing teachers unchanged once Balsakhis become part of official policy (or that any changes will not decrease the gains to learning). In fact, regular teachers subsequently started to perceive Balsakhis as a threat and protested against the programme, resulting in its closure in Mumbai municipal schools in 2003 (Pallavi 2005). Thus in this particular case, the existing political concerns within the system—presumably concerns about employment and professional identity—would need to be taken into account while designing the test intervention and extrapolating from any findings.

We now turn to the DHR case study, where the test context lies outside the target one, sidestepping any immediate problems of this kind, but calling for additional assumptions to arrive at policy conclusions.

4.2 The camera intervention

DHR's main finding is that absenteeism rates fell significantly in treatment schools relative to control schools. Since the final outcome of interest is students' learning, they demonstrate that test scores also witnessed a corresponding increase. That is, teachers maintained sufficient 'effort' while in school, enabling the increase in their presence to translate into more learning.

DHR's main conclusions are that

> ...the barriers currently preventing teachers from attending school regularly (e.g., distance, other activities) are not insurmountable. Given political will, it

> is possible that solutions to the absence problem could be found in government schools as well.
>
> ...para-teachers can be effective. If implementing monitoring within the government system turns out to be impossible, our results provide support for the policy of increasing teaching staff through the hiring of para-teachers.
>
> (DHR, p. 1276)

As noted above, the test population is a group of teachers in Non Formal Education centers run by an NGO in western India, while the policy target are teachers in (formal) government schools. Their main policy recommendation—for para teachers in the government school system—rests on three assumptions. The first assumption is implicit and relates directly to our argument about choice of interventions: the assumption that para teachers, potentially coupled with absenteeism-checking mechanisms, are better than existing government school teachers. As with BCDL, this is a normative assumption upon which the policy recommendations of the study crucially hinge, yet DHR do not provide direct or indirect evidence in support of it, since the study neither deals with government school teachers nor compares them with the test population of NFE teachers. In order to accept this assumption, we must first accept that candidate policies are to be drawn from a restricted pool that excludes, for example, changes to teacher training, school curricula and various corresponding institutions.

The second and third assumptions correspond to the twin forms of extrapolation referred to earlier: that similar results would maintain in the distinctly different policy target context of government schools, and that it would be feasible to implement similar schemes to check the absenteeism behavior of government school para teachers.

DHR also present a theoretical model as the causal explanation for why teacher absenteeism fell in response to the intervention. Any objections to their policy recommendations—including our criticism of their restricted intervention choice—would be less relevant if this explanation could be used to (a) predict how government school teachers will respond to a similar absenteeism-curbing policy; and (b) compare para teachers with government school teachers in terms of children's learning net of any absenteeism differences. This brings us to the second part of our argument that focuses on the role of theoretical explanations and their ability to strengthen extrapolation.

5 The role of theory

In his celebrated critique of RCTs and econometric practice more generally, Deaton (2010a, p. 452) appeals for better causal explanations, stating that '...we are unlikely to banish poverty in the modern world by [randomised controlled] trials alone, unless those trials are guided by and contribute to theoretical understanding.' Related to this is Harrison's (2013) critique of current practice in field experiments, who argues that causal evidence of a change in average outcomes, while far from

being the only interesting kind of inference, needs economic theory to explain *why* human beings responded in a certain way to some policy.

Do or can theoretical models fulfill this role, and equip us to make plausible guesses about how a given intervention will play out in the target population and how it will compare with others? Our claim is that in general the answer is no when the aim is to generalise from RCT results to policy advice for developing country public systems. The answer cannot be definitive because the notion of a theory or theoretical model can, in principle, always be made sufficiently rich so as to *always* succeed in gleaning a useful nugget of knowledge from *any* piece of evidence, such as from an internally-valid RCT. Instead, our task is to examine this question with regard to theory as it currently exists within the discipline of economics, in the form of mathematico-deductive theoretical models.

5.1 Models in economics

Heckman's (2005, p. 2) explanation of causality introduces models as '...a set of possible counterfactual worlds constructed under some rules', which as he explains, are an essential prerequisite for conceptualizing causality. Models are the opposite of a black-box; replacing unknown relationships with known, or at least hypothesised connections between different variables and processes that influence one another. Given that the usefulness of an RCT rests on the continued validity of its findings over time, in different places, or with different people, models could aid both in conceptualizing and in assessing this generalizability. They might help think through the preconditions necessary for a certain bit of extrapolation, or how an intervention might need to be amended or supplemented under different circumstances. This view is echoed by Deaton (2010a), who, like Harrison (2013), focuses not so much on an exposition of the relationship between models and causality so much as on the need for a causal understanding for producing credible knowledge and drawing policy-relevant conclusions. The emphasis here is on constructing theoretical mechanisms through which successive research studies can progressively refine interventions in Popperian fashion.

Such models are usually mathematico-deductive in nature, and consist of rules by which agents interact with other and with institutions (e.g. a company or NGO), typically governed by a set of constraints. Within a model, agents are assigned behavioural rules, which for individuals usually entail complete rationality or carefully defined departures from this.[4] The rules by which agents interact form 'logically consistent systems within which hypothetical "thought experiments" can be conducted' (Heckman 2000, p. 46), and they constitute the bulk of theory in most economics practice. The outcomes of these interactions—and predictions of the model—come about by using comparative statics to analyze the collective interplay of different actors and constraints in the model. Such a model can also be used to define a structural econometric model, to help analyze data in a way that

[4] Rationality has been criticised for being unrealistic (Sen 1977), but also for being at odds with the 'ecological rationality' that exists in societies with non-liberal economic relationships (Davis 2013).

makes explicit the distinction between association and causation (Moneta and Russo 2014; Deaton 2010) as in DHR's model.

5.2 Models for RCTs

A priori, there is no binding link between a given methodology and the type of model chosen to explain the causal connections this methodology seeks to uncover. A causal explanation for the findings of an RCT could thus be built, in principal, using a sufficiently detailed model. Such a model could help in analyzing the generalizability of any findings by capturing relevant preconditions and contextual factors in test and target populations.

Rubinstein (2006) states that as a first step towards hypothesizing useful explanations, a model should yield conclusions that are consistent with observed phenomenon. Sugden (2009) goes further, arguing that while models aim to provide explanations about the real world, model-builders usually hesitate to claim this explicitly. Indeed 'within economics, *explicit discussion of the relationship between a model and the corresponding real-world phenomena is not required*' (p. 9, author's emphasis). Instead, Sugden suggests, models can be thought of as 'credible worlds', i.e. fully functional constructions in themselves and not simplified versions of the real world. The model builder, in such cases, generally leaves unstated the inferential leap from model to causal explanation for real world phenomena. How does this view apply to the model offered by DHR?

As described in Sect. 2.2, DHR's model consists of a principal-agent setup where teachers maximise their utility and are imperfectly monitored by the NGO. In this mathematico-deductive model world, control group teachers face a simple tradeoff between a day's leisure and the chance of being rebuked if caught while absent. Treatment group teachers face a dynamic incentive on account of the per-day payment they receive (for every day they are present beyond an initial ten days in the month), and they decide on whether to attend by weighing the benefit of an additional day's salary against the value of leisure and the chance of being rebuked or fired. For all teachers, the value of leisure is hypothesised net of any intrinsic reward from teaching. Teachers were given an achievement test as part of the study, and their test score is assumed to proxy their intrinsic reward from teaching 'to control for the fact that teachers with higher scores may be more diligent and thus may work more often.' (p. 1258). The model does not consider systemic attributes of the kind government schools might be expected to have, including higher levels of administration, teacher deployment and in-service training, or potential variation in the profiles of children being taught.

5.2.1 Credible worlds

Under Sugden's credible-worlds view, this model must in fact constitute a self-contained universe with NFE teachers and the NGO as the main actors. Sugden also states that most models in economics are not accompanied by an explicit declaration or explanation pointing out the correspondence with the real world we might believe they are intended to represent.

DHR's model conforms to this: the authors do not declare how to infer from their model to the real world; either that of the NFE schools, or more importantly government schools, the policy focus of their analysis. Indeed, their model does not consider aspects or actors beyond teachers and a monitoring authority. If the model is intended as a bridge between context-specific results and the target public system, this bridge does not attempt to describe that target system. Viewed in terms of the structural versus reduced-form debate (Chetty 2009; Heckman 2010), accounting for the public system here would be akin to a 'super-structural' model. This would describe not just the variables that jointly determine absenteeism behavior in the test context—a simple structural model like the one provided—but also all relevant variables in the target public system.

To be clear, this point is distinct from the idea that DHR's model employs unrealistic assumptions. Economic models do so in general, and as Mäki (2009) explains, it is mistaken to argue that such lack of realism makes models any less useful. Models attempt to isolate hypothesised mechanisms, and their usefulness is largely independent of the realism of their assumptions. Instead, my suggestion is that DHR's model does not consider the target (public) system at all. And, that since this is due in part to the complexity of such systems relative to the model-building tools employed in economics, this inability will likely be a characteristic of models in general.

But, this criticism could be weakened if the causal mechanism inside the model-world also operates in the target system. A similar treatment effect would manifest in the target system provided that target is free of interfering factors. In the case of DHR, the policy recommendation of incentive-based salaries for teachers in the government education system might yield results similar to the tested intervention provided there are no new variables in this larger system which influence absenteeism. Nancy Cartwright (2009) makes this condition precise.

5.2.2 Isolating mechanisms

Cartwright suggests that models attempt to isolate key mechanisms that do actually function in the real world, and that doing so lets us study the effects of a 'capacity'. In DHR's model, the capacity of interest is the incentive-payment scheme, and their model offers a mechanism for its effect on absenteeism. For the model's explanation to be useful, it is necessary to eliminate all confounding factors or assume that they are orthogonal to the treatment mechanism. So, since the DHR model ignores teacher training and governance structures (say), these must not be factors that shape the link between incentive pay and absenteeism, neither with these NFE teachers, nor—because the model must extrapolate to the government schools they offer policy prescriptions for—for government school teachers.

But to isolate key mechanisms, Cartwright explains, economic models tend to invoke several structural assumptions. These are needed because economics lacks a useful body of general principles (unlike the physical sciences), muddying the task of extrapolation because valid inference from model to real world depends on the validity of these structural assumptions. In DHR's model, the primary structural assumptions are utility maximizing behavior by the teachers, and an absence of any

influences on their behavior other than those included in the model. Extrapolating to government schools requires believing that this behavior stays invariant for teachers in government schools, and moreover, that its ultimate influence on teaching and thus learning remains invariant to any confounding influences in this alternative context. The latter include teachers' characteristics and those of the children they teach, school organization and administrative structures.

Similar assumptions would need to be made for any RCT that focuses on individuals but whose target in terms of policy advice through extrapolation is a larger system. Here, the theoretical model must not only set out the key structural assumptions that explain causal mechanisms within the experimental context, but also any additional structural assumptions that are important characteristics of the target system. Since the experimental context is, by design and feasibility restrictions, often very different from the target context, this places a substantial onus on the model. DHR's model effectively illustrates the challenge of doing this using the tools of theoretical modeling as they currently exist in economics. Here and more generally, it is difficult to account for the structural assumptions and characteristics of a target public system, or alternatively to build a 'credible world' representation of it, where inference from model-plus-experimental results to the real world can be justified through explicit, plausible assumptions.

6 Conclusion

I have argued why RCTs should not be granted the status of preferred method for evaluating policies aimed at improving public systems in developing countries. The set of interventions that are amenable to randomization is a restricted set owing to the need for feasibility, both in terms of scale and time. When a public system is the target of a policy evaluation exercise, this restriction means that the intervention actually tested tends to ignore how the system is organised, and is implemented only for the last tier of workers. The necessity of choosing from a restricted set of interventions also implies that normative judgements must be made regarding the choice of candidate intervention, and these judgements are seldom made evident in a discussion of the findings.

Two types of extrapolation are required for effective policy to emerge from research findings: from the test to target—public system—context, and from the tested intervention to the version of it that actually follows in policy. I have argued as to the challenges of doing so based on results from RCTs, and have illustrated how important yet unstated and untestable assumptions are needed to believe that a similar causal mechanism will still operate once other layers of a system become involved.

Clearly, these criticisms could be weakened or alleviated if sufficiently detailed causal explanations existed. The process and validity of extrapolating from test to target context could be appraised by comparing these twin contexts from the perspective of a known causal mechanism. And, once this mechanism is identified, it would be possible to assess an allied set of interventions known to have certain features in common with the intervention actually tested, thereby easing the

problems of both normative judgement in choosing an intervention and of the constraints imposed by feasibility.

In response, I have explained how the contemporary form of causal explanations in economics does not permit for modeling complex systems with multiple constraints, incentives, and agents. Current best practice, as illustrated by DHR's theoretical model, usually explains select features of observed behavior through a framework of utility-maximization or careful deviations therefrom. At best, such practice can provide a causal explanation for what is observed in a subset of a public system, but it does not discuss the validity of extrapolating to other parts of the system, nor how the hypothesised 'capacities' might operate in the target context. As a result, the problems associated with extrapolating from experimental results continue to be serious concerns, and particularly so when public systems are the target of that extrapolation.

In principle, this criticism can be levelled against any attempt to design policy for a public system based on findings from a small, potentially different test context. What makes public systems in developing countries stand out is that these are often still in the process of being established, and they might lack basic levels of functional competence. Critically, fundamental questions about their role and organisation might still be open to debate: the role of private delivery, how to organise governance structures, and the numbers, type, and qualifications of personnel needed. This undecidedness puts a distinct onus on research leading to policy prescriptions, as we have critically analysed in the case of contractual teachers and their role within a government school system. Any claim to the preeminence of a methodology must reflect its ability to evaluate a wide range of policies, and provide a valid conceptual if not empirical basis to extrapolate from test to target. It is in this regard that the claim towards RCTs as preferred method is problematic.

Acknowledgments I am grateful to Shaun Hargreaves-Heap, Ragupathy Venkatachalam, and three anonymous referees for their comments. All errors are my own.

References

Armstrong, D. (2006). Trial and error. *Forbes*. http://www.forbes.com/forbes/2006/0619/128.html. Accessed 27 Oct 2014.

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: evidence from two randomized experiments in India. *The Quarterly Journal of Economics*, *122*(3), 1235–1264.

Banerjee, A. V., & Duflo, E. (2011). *Poor economics: A radical rethinking of the way to fight global poverty*. London: Penguin.

Barrett, C. B., & Carter, M. R. (2014). A retreat from radical skepticism: rebalancing theory, observational data, and randomization in development economics. In D. L. Teele (Ed.) Field

experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences (pp. 58–77). Yale University Press.
Black, R. E., Morris, S. S., & Bryce, J. (2003). Where and why are 10 million children dying every year? *Lancet*, *361*, 2226–2234.
Blimpo, M. P. (2014). Team incentives for education in developing countries: a randomized field experiment in Benin. *American Economic Journal: Applied Economics*, *6*(4), 90–109.
Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties*, *2*, 11–20.
Cartwright, N. (2009). If no capacities then no credible worlds. But can models reveal capacities? *Erkenntnis*, *70*, 45–58.
Cartwright, N. (2011). A philosopher's view of the long road from RCTs to effectiveness. *Lancet*, *377*, 1400–1401.
Chetty, R. (2009). Sufficient statistics for welfare analysis: a bridge between structural and reduced-form methods. *Annual Review of Economics*, *1*, 451–488.
Davis, J. B. (2013). Economics imperialism under the impact of psychology: the case of behavioral development economics. *Oeconomia*, *3*(1), 119–138.
De, A., Khera, R., Samson, M., & Kumar, A. S. (2011). *PROBE revisited: A report on elementary education in India*. New Delhi: Oxford University Press.
Deaton, A. (2010a). Instruments, randomization, and learning about development. *Journal of Economic Literature*, *48*, 424–455.
Deaton, A. (2010b). Understanding the mechanisms of economic development. *The Journal of Economic Perspectives*, *24*(3), 3–16.
Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives work: getting teachers to come to school. *American Economic Review*, *102*(4), 1241–1278.
Favereau, J. (2014). PhD summary: "The J-PAL's experimental approach in development economics: an epistemological turn?". *Erasmus Journal for Philosophy and Economics*, *7*(2), 177–180.
Glewwe, P., Kremer, M., Moulin, S., & Zitzewitz, E. (2004). Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. *Journal of Development Economics*, *74*(1), 251–268.
Government of India (1998). *National Policy on Education 1986 (as modified in 1992)*. New Delhi: Ministry of Human Resource Development (Revised 1992).
Govinda, R., & Josephine, Y. (2004). *Para teachers in India: A review*. New Delhi: National Institute of Educational Planning and Administration.
Grüne-Yanoff, T., & Hertwig, R. (2015). Nudge versus boost: how coherent are policy and theory? *Minds and Machines*. doi:10.1007/s11023-015-9367-9.
Guerrero, G., Leon, J., Zapata, M., Sugimaru, C., & Cueto, S. (2012). *What works to improve teacher attendance in developing countries? A systematic review*. EPPI-Centre: Social Science Research Unit, Institute of Education, University of London.
Halperin, R., & Ratteree, B. (2003). Where have all the teachers gone? The silent crisis. *Prospects*, *33*(2), 133–138.
Harrison, G. W. (2013). Field experiments and methodological intolerance. *Journal of Economic Methodology*, *20*(2), 103–117.
Haynes, L., Service, O., Goldacre, B. and Torgenson, D. (2012). (2012). *Test, learn, adapt: Developing public policy with randomised controlled trials*. London: UK Cabinet Office and Behavioural Insights Team.
Heckman, J. J. (2000). Causal parameters and policy analysis in economics: a twentieth century retrospective. *The Quarterly Journal of Economics*, *115*(1), 45–97.
Heckman, J. J. (2005). The scientific model of causality. *Sociological Methodology*, *35*, 1–97.
Heckman, J. J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic Literature*, *48*(2), 356–398.
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.
Kingdon, G. G., & Sipahimalani-Rao, V. (2010). Para-teachers in India: status and impact. *Economic and Political weekly*, *45*(12), 59–67.
Kremer, M., Brannen, C., & Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science*, *340*(6130), 297–300.
Kremer, M., Chaudhury, N., Rogers, F. H., Muralidharan, K., & Hammer, J. (2005). Teacher absence in India: a snapshot. *Journal of the European Economic Association*, *3*(2), 658–667.
Kumar, K., Priyam, M., & Saxena, S. (2001). The trouble with 'para-teachers'. *Frontline*, *18*(22), 93–94.

Mäki, U. (2009). Realistic realism about unrealistic models. In H. Kincaid & D. Ross (Eds.), *The Oxford handbook of philosophy of economics* (pp. 68–98). New York: Oxford University Press.
Miguel, E., & Kremer, M. (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, *72*(1), 159–217.
Moneta, A., & Russo, F. (2014). Causal models and evidential pluralism in econometrics. *Journal of Economic Methodology*, *21*(1), 54–76.
Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: experimental evidence from India. *The Journal of Political Economy*, *119*(1), 39–77.
NCERT. (2006). *National Curriculum Framework 2005*. New Delhi: National Council for Educational Research and Training.
(2014). *Elementary education in India: Progress towards UEE: Flash Statistics 2013–14*. New Delhi: National University of Educational Planning and Administration.
Pallavi (2005). Forms and discourse of NGO/government partnerships: the case study of Pratham and Childline India Foundation. Master's thesis, Institute of Social Studies, The Hague.
Reddy, S. (2012). Randomise this! On poor economics. *Review of Agrarian Studies*, *2*(2), 60–73.
Rodrik, D. (2009). The new development economics: ee shall experiment, but how shall we learn? In J. Cohen & W. Easterly (Eds.) What works in development? Thinking big and thinking small (pp. 24–47). Brookings Institution Press.
Rubinstein, A. (2006). Dilemmas of an economic theorist. *Econometrica*, *74*(4), 865–883.
Sen, A. K. (1977). Rational fools: a critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs*, *6*(4), 317–344.
Stokes, S. C. (2014). A defense of observational research. In D. L. Teele (Ed.) Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences (pp. 33–57). Yale University Press.
Sugden, R. (2009). Credible worlds, capacities and mechanisms. *Erkenntnis*, *70*, 3–27.
UNESCO. (2005). *Global Monitoring Report 2005: education for all: the quality imperative*. Scientific and Cultural Organization: United Nations Educational.
World Bank (2003). *World Development Report 2004: Making Services Work for Poor People*. Oxford University Press for the World Bank.
Worrall, J. (2007). Why there's no cause to randomize. *The British Journal for the Philosophy of Science*, *58*(3), 451–488.