The importance of specifying and studying Causal Mechanisms in School-based Randomised Controlled Trials: Lessons from two Studies of Cross-age Peer Tutoring

Stephen P. Morris*

Stephen is Professor of Evaluation at the Policy Evaluation Research Unit, Department of Sociology, Manchester Metropolitan University

Department of Sociology,
Manchester Metropolitan University
Room 414, Geoffrey Manton Building
4 Rosamond Street West
Manchester
M15 6LL
Tel. +44 (0) 7730 234178
Email. s.morris@mmu.ac.uk

Dr Triin Edovald

Triin is Principal Researcher at the Innovation Growth Lab, Nesta

Innovation Growth Lab at Nesta
1 Plough Place
London
EC4A 1DE
Tel. +44 (0) 7438 2558
Email. triin.edovald@nesta.org.uk

Cheryl Lloyd

Cheryl is Research and Policy Manager at the Nuffield Foundation

Nuffield Foundation
28 Bedford Square
London
WC1B 3JS
Tel. +44 (0) 20 7681 9595
Email. clloyd@nuffieldfoundation.org

Dr Zsolt Kiss

Zsolt is a Chief Behavioural Data Scientist at ZK Analytics

ZK Analytics
Oxford Centre for Innovation
New Road
OX1 1BY
Tel. +44 (0) 7454572266
Email. zsolt.kiss@zkanalytics.com

* Corresponding author. Email: s.morris@mmu.ac.uk

The importance of specifying and studying Causal Mechanisms in School-based Randomised Controlled Trials: Lessons from two Studies of Cross-age Peer Tutoring

Based on the experience of evaluating two cross-age peer-tutoring interventions, we argue that researchers need to pay greater attention to causal mechanisms within the context of school-based randomised controlled trials. Without studying mechanisms researchers are less able to explain the underlying causal processes that give rise to results from randomised controlled trials. Studying implementation fidelity is necessary but not sufficient for causal explanation; the study of causal mechanisms through the application of mixed methods is also required. Due to the increasingly complicated nature of many classroom-based innovations that are subject to evaluation, and the potentially distal nature of hypothesised effects, particularly on attainment, programme theory and articulation of mechanisms are essential in enhancing causal explanation and promoting the accumulation of knowledge of what works and why in classroom settings.

Keywords: randomised controlled trials, mechanisms, causal explanation, process evaluation, programme theory

Introduction

Through the activities of the Education Endowment Foundation (EEF) the number of randomised controlled trials (RCTs) in education in England has increased appreciably in recent years. So much so that attention has turned to the limitations of RCTs and the development of implementation and process evaluation (IPE) methods to enhance causal explanation (Humphrey et al., 2016a, 2016b; Lendrum & Humphrey, 2012). This paper aims to contribute to this debate through stressing the importance of studying 'mechanisms' and therefore the importance of programme theory[1] in the

[1] We follow Funnell and Rogers (2011) and define programme theory as comprising two components, a theory of action and a theory of change. From our perspective, the former relates to the study of implementation and fidelity, while consideration of the latter in the

context of school-based RCTs. Understanding the processes through which effects are produced is essential if trials are not only to identify causal effects but explain their presence or absence.

This paper makes the case for the study of mechanisms within RCTs. It is not concerned with a detailed discussion of research design or methodology, although we provide a brief overview of recent and more commonly encountered approaches to the study of mechanisms to demonstrate their existence and potential utility. We acknowledge that the study of mechanisms, and indeed intervention implementation, is challenging. But, as we will demonstrate, there is much to be gained from trialists undertaking such investigations despite these challenges.

We argue that the study of mechanisms within trials, alongside the evaluation of implementation fidelity contributes to the capacity to explain causal effects. Regardless of whether RCTs find interventions to be effective or ineffective, it is important to be able to explain why[2]. If researchers are limited in their ability to explain causal effects it becomes difficult to identify how weak but potentially important interventions might be improved. It is also more difficult to accumulate knowledge of what works both theoretically and practically, and practitioners will finder it harder to judge whether interventions will work in their particular circumstances.

Many school-based teaching interventions are complex and multifaceted, with intended outcomes that are distal rather than proximal to the intervention. These features of school-based interventions highlight the importance of having an explicit theory of how the intervention gives rise to anticipated effects; the processes or

---

context of an RCT provides a theory of how an intervention's proximal outcomes are linked to more distal effects.

[2] Raudenbush (2008) points out the importance of being able to explain how a school-based intervention works in both cases where the evidence points to the intervention being effective and where there is an absence of effectiveness. The requirement for enhanced causal explanation is particularly strong in the case of the latter.

mechanisms that lie on the causal pathway between the intervention and the ultimate outcome.

Drawing on two examples of school-based RCTs, we argue that the study of mechanisms would have helped researchers explain the apparent failure to find an impact on attainment in both cases, and enabled results from both evaluations to make a greater contribution of the future development of peer-tutoring interventions. Unfortunately, in neither trial were mechanisms the focus of the research. This omission is not surprising, as the study of mechanisms is only rarely addressed within the context of school-based RCTs. Despite the widely acknowledged importance of mechanisms in understanding causal processes both in the social sciences in general (Goldthorpe, 2001; Hedstrom & Ylikoski, 2010) and in evaluation research in particular (Pawson & Tilley, 1997; Weiss, 1997), it is important to acknowledge at the outset that there is considerable disagreement in the literature over how mechanisms are to be understood, defined and tested empirically (Gerring, 2010). These difficulties may explain the general absence of attempts to studying mechanisms in the context of RCTs in education. Despite these challenges, we argue that researchers conducting school-based RCTs need to adopt mechanistic forms of explanation, and develop methods that can explore mechanisms in useful ways. The lack of attention paid to mechanisms leads to considerable ambiguity in attempts to explain many trial results.

We commence by providing a brief discussion of what is meant by mechanisms, drawing primarily on the programme evaluation literature and highlight definitions consistent with those suggested by authors working in the counterfactual tradition[3]. We then focus attention on recent discussion in the literature concerning some of the more common approaches to studying mechanisms currently seen as promising among

[3] We locate RCTs within the more general counterfactual approach to causation

researchers. We briefly introduce the two trials that represent our case studies, describing their design and the results, highlighting the design weaknesses that frustrated attempts to explain why neither intervention was found to be effective. We then move on to discuss in broad terms the study of mechanisms within pilot, efficacy and effectiveness trials and examine how consideration of mechanisms might have improved causal explanation in the case of the two trials discussed.

What are mechanisms?

Combining insights from mechanistic approaches to causal explanation with RCTs has been controversial. For example, those who espouse a generative approach[4] to causation, within which mechanisms are a central concept, have gone as far as to suggest that consideration of generative processes acts not as a "complement" to the types of evidence emerging from RCTs but instead as a "corrective" (Goldthorpe, 2001, p. 9). The realist evaluation tradition, with its emphasis on context-mechanism-outcome configurations, either flatly rejects or largely discounts RCTs as a means of generating useful knowledge (Pawson & Tilley, 1997). Others have raised concerns over the weaknesses of RCTs in relation to external validity stemming in part from an inability to explain causal effects in these terms (Cartwright & Hardie, 2012).

We argue, however, that there is no inherent contradiction in framing causal processes in terms of mechanisms within the context of RCTs in education. Moreover, that there is a long tradition of testing for the presence of mechanisms within RCTs going back, at least, to the work of Baron and Kenny (1986). Although Baron and Kenny and those inspired by their work rely exclusively on quantitative measures of

[4] Goldthorpe (2001, p. 9) defines the concept of generative causation as "some process existing in time and space, even if not perhaps directly observable, that actually generates the causal effect of X on Y and, in do doing, produces the statistical relationship that is empirically in evidence". Our view is that the notion of generative causation is consistent with the concept of a mechanism as discussed in this paper.

mechanisms (in our view a weakness), as Morgan and Winship (2007, p. 219) argue "there is no incompatibility between causal mechanisms and counterfactual thinking".

RCTs, although providing secure evidence of the existence of causal effects have obvious limitations in this regard. As Shadish, Cook, and Campbell (2002, p. 9) note in the introductory chapter to their classic text on experimental design:

> The unique strength of experimentation is in describing the consequences attributable to deliberately varying a treatment…..in contrast, experiments do less well in clarifying the mechanisms through which and the conditions under which the causal relationship holds

Indeed, Morgan and Winship (2007, p. 219) commenting more broadly on the capacity of counterfactual approaches to provide a sufficiently satisfactory account of causal effects note:

> It is widely recognized that a consistent estimate of a counterfactually defined causal effect of D on Y may not qualify as a sufficiently deep causal account of how D effects Y

The increased use of RCTs to improve knowledge of effective teaching is a hugely positive development. Such trials, with their focus on causal description offer the prospect of gaining secure knowledge of the effects of innovations in teaching practice on pupil attainment. The recent renewed emphasis on causal explanation with its particular focus on implementation fidelity[5] is also welcomed. The extent of learning from RCTs will, however, be limited if researchers do not pay attention to the mechanisms that give rise to the effects identified in the statistical analysis of trial data.

[5] We understand implementation fidelity to refer to the extent to which the intervention is delivered as it was intended (Gearing et al., 2011).

The term mechanism has been used to refer to a variety of different phenomenon within the social scientific literature. What is clear in all these discussions is that mechanisms are "an irreducibly causal notion" (Hedstrom & Ylikoski, 2010, p. 50). But the range of uses of the term and competing definitions has led to difficulties applying the concept in practical research (Gerring, 2010; Peterson, 2016).

As a first step, however, it is helpful to distinguish mechanisms from features of the intervention under consideration. Often terms such as 'intervention mechanism' are used suggesting that a mechanism is some aspect or feature of the intervention. The evaluation literature is helpful in this regard. As Weiss (1997, p. 48) makes clear:

> An evaluation that attempts to track the theoretical underpinnings of the program has to devise ways to define and measure the psychosocial, physiological, economic, sociological, organizational, or other processes that intervene between exposure to the program and participant outcomes

And again (Weiss, 1997, p. 46):

> The mechanism of change is not the program activities per se but the response that the activities generate

In the context of RCTs in education, Peterson (2016, p. 304) offers the following definition, which is not inconsistent with that of Weiss:

> Remaining within a counterfactual paradigm, a workable definition of key mechanisms in education is that mechanisms are the proximate and most significant factor impacting on change in learning behaviours or understanding.

These definitions suggest that the study of mechanisms is a matter of prioritising the most important processes operating between the intervention and more distal outcomes such as attainment. A broader point, however, is that a mechanism is something that connects the new or reformed teaching practice to the final outcome. In

the context of school-based interventions, mechanisms are the responses of pupils, teachers and other staff to the resources and information made available through the intervention. The mechanism is not a feature of the intervention, nor some form of process embedded within it.

How does the study of mechanisms relate to the evaluation of implementation fidelity? Although it is necessary that trials embrace the study of implementation fidelity (Lendrum & Humphrey, 2012) such study is not sufficient to explain the presence or absence of causal effects. As Raudenbush (2008) points out, null findings can be explained by theory failure and/or implementation failure. The study of implementation fidelity addresses the latter of these explanations but not the former.

Another important feature of mechanisms within the evaluation literature is the link with context and the way in which mechanisms are understood as inherently contingent phenomena. One of the main contributions of the realist approach to evaluation is to stress the sensitivity of mechanisms to context (Dalkin, Greenhalgh, Jones, Cunningham, & Lhussier, 2015). This implies the necessity to study mechanisms and their interaction with context in both efficacy and effectiveness trials due to the potential for contextual factors to differ markedly. The consideration of mechanisms within an RCT also implicitly acknowledges the expectation that results from testing an intervention will vary from setting to setting. Thus understanding what features of settings explain such variation is of critical importance.

How have researchers sought to address mechanisms in the context of RCTs?

The study of mechanisms within RCTs is often framed by an acknowledgment of the complexity of many social interventions, including school-based interventions, and the requirement to explain causal effects through unpacking the so-called black box.

Due to the interplay between human agency and social structure the processes through which effects are generated cannot be simply assumed (Craig et al., 2008).

Although initially framed around a call for mixed method process evaluation, the challenge of explaining the effects of *complex* interventions within the context of school-based RCTs has been acknowledged for some time by researchers studying health interventions implemented in school settings (Oakley, Strange, Bonell, Allen, & Stephenson, 2008). More recently, researchers have advocated realist-RCTs, comprising a focus on mid-level theory (or programme theory) and context-mechanism-outcome configurations as a means of addressing the classic black box critique of RCTs testing complex social interventions (Bonell, Fletcher, Morton, Lorenc, & Moore, 2012, 2013). These attempts at enhancing causal leverage within trials have attracted criticism from those alarmed at the prospect of epistemological confusion arising from the attempt to synthesise RCTs with 'realist' methods (Marchal et al., 2013). Nonetheless, those advocating this methodological synthesis continue to develop their approach (Jamal et al., 2015). However, it isn't entirely clear how those proposing such a synthesis define the concept of a mechanism, and whether their understanding is consistent with that of the realists. This potential incompatibility at a conceptual level does threaten the prospects for a satisfactory synthesis to emerge, and further, that such attempts at a synthesis maybe an unnecessary distraction from the task of enhancing causal explanation within RCTs more generally.

In addressing the challenge of causal explanation within RCTs in political science and economics, researchers have recently returned to the pioneering work of Baron and Kenny (1986) and their moderator/mediator analysis (Imai, Keele, & Tingley, 2010; Imai, Keele, Tingley, & Yamamoto, 2011; Keele, Tingley, & Yamamoto, 2015). Indeed, those espousing realist RCTs also suggest incorporating

such mediator/moderator analysis, alongside other approaches, in their work (Jamal et al., 2015).

Within such a framework, moderator variables, observed prior to randomisation, can be viewed as capturing aspects of context. These variables can be used in statistical analyses to explore how estimated causal effects vary by moderator or subgroup[6]. However, moderator analysis suffers from a number of drawbacks. Strong prior knowledge is required to identify the right moderator variables at baseline, sample size requirements are often substantial, and inflation of Type II statistical error rates are a concern with multiple interaction tests required to explore a full range of relevant hypotheses. Moreover, contextual factors have to be operationalised as quantitative variables and capable of being introduced into the standard analytical framework. These weaknesses suggest a role for qualitative research in strengthening such analyses.

Mediating variables, on the other hand, can be interpreted directly as causal mechanisms consistent with the definition of mechanisms offered by Weiss; in fact Weiss makes this very point in her 1997 paper when discussing Baron and Kenny (1986) (Weiss, 1997). These ideas have been extended in recent discussion where researchers have proposed mechanism experiments (Imai, Tingley, & Yamamoto, 2013; Ludwig, Kling, & Mullainathan, 2011; Peterson, 2016).

The challenge of obtaining unbiased estimates of mediating variables is that the mechanisms themselves are not randomly assigned and are only indirectly manipulated in programme evaluations. Identification of causal effects relies on a range of restrictive assumptions. Mechanism experiments attempt to address this shortcoming through

[6] Jamal et al. (2015, p. 6) use the phrase ‘contextual contingencies’ when discussing moderators in the context of realist RCTs.

manipulating a hypothesised causal mechanism directly[7]. As yet, however, such approaches are rarely found in the literature, particularly in education. Moreover, it is not clear that policymakers will be willing to fund such trials when their concerns typically focus on the effectiveness of a specified intervention.

In examining current practices with regard to the identification of causal mechanisms it is worth considering a parallel debate within the political sciences, where researchers have discussed the concept of causal process observations. Collier, Brady, and Seawright (2010b) argue for an enhanced role for qualitative forms of causal inference. The importance of qualitative research in enhancing causal leverage has also been asserted by education researchers (Maxwell, 2012). Though their arguments have focused on combining qualitative insights with analyses from quantitative observational studies and natural experiments, Collier, Brady and Seawright's proposed framework would appear promising from the perspective of RCTs (Dunning, 2008; Paluck, 2010).

Collier et al. (2010b) make the distinction between two forms of observation - data set observations (DSOs) and causal process observations (CPOs). In the context of a trial, DSOs are the data generated through the process of sampling cases (pupils, classes or schools), randomly allocating the sample to treatments and attaching quantitative measures on the dependent variable to these data (usually measures of attainment). By contrast, a CPO is defined by Collier, Brady, and Seawright (2010a, p.

[7] Ludwig et al. (2011) illustrate the concept of a mechanism experiment by contrasting it with a traditional policy experiment. Ludwig et al. (2011) use as an example testing the effectiveness of the theory that if policy pays more attention to tackling minor low-level crime this reduces the growth of more serious offences because it sends a message that crime generally is not tolerated. This is the so-called broken windows policing strategy. The traditional policy experiment would randomly assign areas to the 'broken window' strategy comparing outcomes in these treatment areas with those in a control group sometime after the commencement of the intervention. By contrast, a mechanism experiment would involve for example selecting areas at random in which steps are immediately taken to clear graffiti, clear rubbish and repair broken windows. Serious crime rates can then subsequently be compared in treated areas with those in non-treated or control areas.

2) as “an insight or piece of data that provides information about context, process, or mechanism, and that contributes distinctively to causal inference”. These insights take the form of detailed knowledge of particular cases as distinct from the kinds of knowledge generated through analyses of DSOs and have their roots in an approach to causal inference known as process tracing (Mahoney, 2010). Usefully, proponents of this approach make the distinction between what they term independent variable CPOs and mechanism CPOs. This is helpful because it clarifies the point that mechanisms are not a feature of an intervention (the independent variable) but the response of participants to that intervention (the mechanism). To date, this discussion has placed much less emphasis on the interactions between mechanisms and context. Examples of precisely how these approaches have been applied within the context of RCTs are also scarce, though the approach would appear to warrant further exploration.

Our purpose is undertaking this brief, though not exhaustive assessment of a range of approaches to studying mechanisms within the context of, or alongside, quantitative approaches to identifying causal relationships, is to show merely that such approaches exists and should be taken seriously by education researchers. Taken together this brief review, suggests that mixed-method approaches, combining both qualitative and quantitative insights are likely to be important in exploring mechanisms within school-based RCT designs.

A tale of two trials

We illustrate the benefits of studying mechanisms in enhancing causal explanation with reference to the limitations of two RCTs of cross-age peer tutoring conducted in English schools (Lloyd, Edovald, et al., 2015; Lloyd, Morris, et al., 2015). Both trials were effectiveness trials. In ideal circumstances, this means that both interventions will have been piloted and shown to be effective in developer-led settings,

with the key features that determine effective practice within each intervention identified. As effectiveness trials, both studies involved the implementation of interventions in settings in which intervention developers/researchers no longer led implementation.

In the case of both interventions the trial results suggest that neither was effective in raising pupil attainment. Unfortunately, it proved difficult to explain why the interventions appeared to be ineffective. This was in part due to inadequacies in the testing implementation fidelity, but also we argue, due to a lack of attention paid to theorising and testing for causal mechanisms.

Commented [SM1]: We have acknowledge the weaknesses in the trial designs here

Cross-age peer tutoring involves older pupils instructing younger pupils. The first intervention, Paired Reading, was tested in English secondary schools and involved Year 7 pupils (aged 11-12 years) being supported by Year 9 pupils (aged 13-14 years) in various reading tasks. The second intervention, Durham Shared Maths, was implemented within English primary school settings and involved Years 5 pupils instructing Year 3 pupils in mathematics. Both interventions were highly structured. The Durham Shared Maths intervention involved two blocks of study comprising 16 weeks activity, whereas the Paired Reading intervention ran for one block of 16 weeks. Both interventions encouraged older pupils to engage in carefully scripted forms of support with their younger counterparts, involving questioning, giving praise and encouragement, and actively reviewing progress.

Prior evidence

Previous research had found that peer-learning interventions were generally effective, and that the approach was favoured as a means of improving attainment in English schools. As Lloyd, Morris, et al. (2015, p. 6) note:

> The evidence for peer tutoring tends to be positive, with reviews showing that peer tutoring is an effective technique for raising attainment in school-aged children, particularly with younger pupils across different subjects including maths, literacy and science – with tutoring in maths being particularly effective

A review from Pennucci and Lemon (2014) considered the effects of both within age-group peer tutoring and cross-age support. Both types of interventions appeared to be effective, but the strength of evidence was greater for within age-group tutoring. Reviews from Britz (1989) and Robinson, Schofield, and Steers-Wentzell (2005) looking specifically at peer tutoring in maths found such interventions to be generally effective.

More recently, the EEF/Sutton Trust concluded in their summary of interventions aimed at raising attainment among pupils, that peer tutoring interventions were potentially effective at raising attainment among tutees and tutors and that pupils from more disadvantaged backgrounds gained most (Higgins et al., 2013).

Finally, a study from Tymms et al. (2011) examined the effects of both peer and cross-age peer tutoring in both English and maths among primary school-aged pupils. The study, undertaken in Fife, Scotland, found that cross-age tutoring had positive effects for both tutors and tutees in both maths and reading.

As a result of this evidence, expectations were that evidence emerging from both trials would be supportive of cross-age peer tutoring and that sufficient development and testing had taken place for effectiveness trials to be justified.

Evaluation designs

The Durham Shared Maths and Paired Reading interventions were evaluated using cluster RCTs combined with process evaluations.

The Paired Reading intervention was carried out in 10 secondary schools, with 60 classes in Year 7 (29 to treatment and 31 to control) and 60 classes within Year 9 randomly allocated to treatment and control conditions within schools (29 to treatment and 31 to control). Pupils in Year 9 treatment classes were then matched with pupils in Year 7 treatment classes to form tutor/tutee pairs – in some cases triplets. The primary outcome for the trial was reading ability measured using the Overall Reading Scale from the New Group Reading Test, implemented at baseline (pre-test) and follow-up (post-test)[8]. Pre-test data were collected from 1,370 Year 7 pupils and 1,366 Year 9 students using a computerised, adaptive, version of the test during September 2013. Similarly, post-test data was collected from 1,306 Year 7 pupils and 1,269 Year 9 pupils in June 2014. Loss to follow-up was minimal at the pupil level and none of the study schools left the trial prior to analysis. Ex-post power tests show the trial was powered to detect standardised mean differences in treatment and control groups at analysis of 0.13 for Year 7 students and 0.11 for Year 9 pupils.

The Durham Shared Maths trial involved the randomisation of whole primary schools rather than classes. In total 82 schools were randomised to treatment and control, 40 to the former, 42 to the latter. Within schools randomised to treatment, pupils from Year 5 were matched to pupils from Year 3 to form tutor/tutee pairs. The primary outcome was students' maths attainment derived from the Interactive Computerised Assessment System (InCAS) module General Mathematics (Merrell & Tymms, 2005) observed at baseline and follow-up. In total, pre-test data were collected from 3,305 pupils in Year 3 and 3,167 Year 5 students over the period September to November 2012. Between baseline and follow-up one school was lost from the

[8] The New Group Reading Test was developed by GL Assessments and the National Foundation for Educational Research (NFER).

treatment arm of the trial and two from the control arm. At post-test data were collected from 2,786 Year 3 students and 2,683 Year 5 pupils. Balance tests conducted on the as randomised and as analysed samples revealed few differences in the mean characteristics of treatment and control schools, and treatment and control pupils at either point. Ex-post power tests revealed the trial was powered to detect standardised mean differences as small as 0.10 of a standard deviation for both Year 3 and 5 pupils.

Both trials incorporated process evaluations. The process evaluation of the Paired Reading trial aimed to address a range of questions around what led schools to take-part, fidelity, sustainability, subjective assessments of effectiveness and formative elements. Thus in terms of its objectives, the process element of this trial should have addressed many of the factors understood from previous research to be important in the assessment of fidelity (Lendrum & Humphrey, 2012), and at least in its intentions reflected EEF's guidance (Humphrey et al., 2016a, 2016b). However, process analysis relied on just eight depth interviews conducted across eight schools: three interviews with senior school leaders, three Year 7 teachers and two Year 9 teachers. The initial ambition of the research team to conduct a survey of teachers focusing on implementation fidelity was thwarted due to resistance from schools. Furthermore, there were no independent assessments made of implementation fidelity on the basis of structured classroom observations.

The Durham Shared Maths process evaluation comprised structured observations of classroom activities and two waves of qualitative depth interviewing. The evaluators conducted 14 depth interviews with teachers as part of the process evaluation. Practitioners responsible for training and supporting teachers were also interviewed in two waves. Two members of the intervention design team contributed data to the study. Structured observations were conducted in two Durham Shared Maths

sessions that focused on: teachers' role and level of guidance, interactions between pupils, pupil understanding of the tasks that required completion, understanding of what was required of pupils at a conceptual level and barriers to implementation.

In the case of neither process evaluation was the definition or empirical study of mechanisms undertaken taken, neither were programme theories developed for either intervention.

Evaluation findings

Table 1 presents findings from both the Durham Shared Maths and Paired Reading trials. Effect sizes[9] estimated from multi-level regression models adjusting for baseline covariates are displayed. The primary outcomes in the case of both trials are the measures of attainment discussed previously. The effect sizes are shown for the primary outcomes only. Table 1 reveals no evidence of impact for either intervention. Effect sizes are close to zero and in some cases negative and do no reach conventional levels of statistical significance.

[INSERT TABLE 1 HERE]

In the Durham Shared Maths study further tests for differences in means by subgroups and for secondary outcomes did not reach statistical significance at conventional levels. A similar picture emerged in further analysis of the Paired Reading trial data.

Turning to the process evaluation results, the researchers who conducted the Durham Shared Maths study concluded that their findings were limited as a result of a

[9] Effect sizes reported are standardised mean differences and therefore in units of standard deviations

lack of systematic, quantitative data on measures of implementation fidelity. Variations in how training for teachers was delivered across sites and that teachers were able to, and did, tailor aspects of the intervention to fit their circumstances were reported. Moreover, that there was some evidence that lower ability pupils were struggling with aspects of the intervention and that teachers did not feel well equipped to support these pupils. On the basis of the evidence that was available, despite its limitations, researchers concluded that Durham Shared Maths was implemented with a reasonably high degree of fidelity. Furthermore, any variations in practice were likely to be within the bounds of that expected in an effectiveness trial.

As discussed, the process evaluation of the Paired Reading trial was hampered by the refusal of schools to engaged with a proposed online survey of the teachers. Moreover, there was no opportunity to undertake structured observations of classroom activities. As a result, conclusions about fidelity and implementation are wholly reliant on teachers' self-reports collected in qualitative interviews.

Bearing these limitations in mind, the process evaluation of Paired Reading found that there had been variation across schools in implementation and delivery of the intervention. Particularly, the amount of time devoted to training pupils, both tutors and tutees. There were also variations in the amount of support to pupils provided by teachers that tended to depend on the initial reading ability of pupils. Pupils received varying levels of treatment depending on the timetabling constraints facing the school. Despite this, researchers concluded that these variations were within the bounds of what might be expected within the context of an effectiveness trial; and moreover, that such variations did not fundamentally alter the extent of pupils' exposure to the intervention.

Is it necessary to study mechanisms in different types of trials?

If the study of mechanisms within school-based RCTs is to be taken seriously we argue that theory needs to be attended to at each stage of the development process. By theory, we refer to programme theory (Funnell & Rogers, 2011) often referred to a mid-level or mid-range theory, which describes both how the intervention is to be implemented and sets out explicitly the intermediate or intervening outcomes that will give rise to the final more distal impact of the intervention. Moreover, a programme theory should describe how outcome chains are affected by contextual factors. A number of causal mechanisms might be hypothesised that give rise to the outcome or causal pathways articulated in a programme theory. The task is then to test for the presence of such mechanisms and to identify which most plausibly account for the trial results.

How then might this discussion relate to different stages in the development and testing of new interventions? The EEF's guidance (Education Endowment Foundation, 2015) specifies three types of trial conducted in sequence in the development and testing of new interventions. These are pilot, efficacy and effectiveness trials, with pilot trials preceding efficacy trials and efficacy trials effectiveness trials in the development and testing process. We consider each of these phases and put forward suggestions for how consideration of causal mechanisms should shape and inform research at each stage. This discussion highlights one of the major problems encountered in the two RCTs discussed; namely, a lack of a clear development process that involved the specification of programme theory, and the identification and testing of causal mechanisms through this process.

Pilot trials – these are early stage studies conducted in a small number of schools with the objective to "develop and refine the approach and test [an intervention's] feasibility" (Education Endowment Foundation, 2015, p. 3). Qualitative

research is expected to be the predominant data collection methodology. The objective is to understand whether a new intervention has potential and the extent to which it is implementation ready.

Our view is that at the commencement of the pilot stage a provisional programme theory should be developed which will be subject to revision in the light of results from the pilot, such that by the end of this stage a theory linking the intervention with the ultimate outcomes it is designed to influence should be capable of clear articulation. The programme theory will enable elements of both implementation and impact to be tested in the subsequent efficacy and effectiveness trials. In the context of a pilot trial, mechanisms can be seen as, ex-ante, the generative processes that are anticipated to give rise to the outcome patterns or causal chains set out in the programme theory. Researchers and developers should be clear about the nature or range of candidate causal mechanisms that are likely to be triggered by the intervention where there are clear competing explanations, as well as which mechanisms are primary or likely to be most important.

Efficacy trials – the objective of an efficacy trial is to explore whether an intervention can work under conditions specified and controlled by the intervention's developers (Education Endowment Foundation, 2015). It is conceived of as a test in conditions that are arranged such that the chances of observing an impact are maximised. Such trials typically comprise both a quantitative impact evaluation as well as mixed methods process evaluation. According to EEF guidance (Education Endowment Foundation, 2015), the role of process evaluation within efficacy trials is to assess elements of effective practice.

Efficacy trials involve identifying whether in statistical terms there appears to be causal relationship between the intervention and the primary outcome of interest. That

is, whether there is a causal relationship to be explained, or whether explanation should focus instead on why the intervention does not appear to have led to the change in outcomes that were anticipated.

Such a process of explanation should draw on programme theory developed at the pilot stage. Implementation will be explored, identifying elements that are considered to be necessary and/or sufficient for the intervention to be effective. Furthermore, both qualitative and quantitative evidence on the degree to which the outcome pathways or outcomes chains posited in the programme theory are observed, and whether the hypothesised causal mechanisms are acting in ways anticipated by the theory. It is important to set out in advance of conducting the efficacy trial how the presence of hypothesised causal mechanisms are to be tested. At this stage null findings in the trial results would suggest exploring whether the programme theory or its implementation is at fault. In either case, it would seem that the intervention should be reformed before being subject to further efficacy testing.

Effectiveness trials test whether the intervention concerned works at scale in circumstances where the developers are no longer solely responsible for implementation and delivery. As with efficacy studies, effectiveness research involves both trial and process evaluation components where the focus of process evaluation is on identifying "challenges and solutions to roll out" (Education Endowment Foundation, 2015, p. 3).

Researchers might assume that the study of mechanisms will not be required by the time an intervention has reached the effectiveness stage. By this point in the development process, researchers and developers should be clear on the causal mechanisms underpinning an intervention and the theory of change upon which it is based. However, aspects of the setting or context in which an intervention is tested are likely to vary between efficacy and effectiveness studies. As discussed, contextual

factors are important due to the contingent nature of causal mechanisms. Studying causal mechanisms in the context of effectiveness studies therefore is likely to involve identifying important features of context and how these interact with mechanisms such that lessons are learnt as to what contextual factors are likely to impede or promote effectiveness in real-world settings.

Discussion

By current prevailing standards within education research both trials discussed in this paper were in quantitative terms well designed and executed. Both had large sample sizes, little or minimal attrition and trial arms were well balanced on baseline covariate values. Appropriate statistical analysis and hypothesis testing was conducted using hierarchical linear regression models and minimum detectable effect sizes were modest, suggesting the trials were sufficiently powered to detect effects of substantive importance.

It is more difficult to judge how far the respective process evaluations provided reliable and valid accounts of implementation fidelity. The evidence that is available suggests that in the case of both interventions there were deviations from planned implementation as specified by developers but that this was not beyond that which might be expected within an effectiveness study. Moreover, it appears that pupils were exposed to both interventions in a manner sufficient to expect average test scores to have improved.

In the light of these conclusions, how might the study of mechanisms within the context of both trials have enhanced causal explanation? We suggest three areas where the development and exploration of programme theory and the study of mechanisms may have contributed to a greater understanding and the enhanced utility of findings. These are: (1) explaining theory failure where implementation fidelity is considered

adequate; (2) directing attention toward the importance of contextual factors; and (3) making a greater contribution to the effective accumulation of knowledge.

Before progressing with this discussion, however, it is worth stressing the importance of the relationship between the study of mechanisms and implementation fidelity. We suggest that they are both necessary but singularly insufficient for causal explanation. Implementation failure can result in the failure of mechanisms to trigger in ways consistent with an intervention's programme theory. Thus in many cases it is possible that supposed theory failure is the result of implementation failure. However, it is also possible that an intervention implemented faithfully can fail due to problems with the underlying theory of change that occurs quite independently. This acknowledgment of the relationship between these two aspects of causal explanation, notwithstanding the inherent difficulties in conducting the such research, underlines the importance of exploring both implementation fidelity and mechanisms in arriving at a satisfactory explanation of causal effects.

Explaining theory failure where implementation fidelity is considered adequate

Given that both trials were well designed and executed, and if findings can be said to show that both interventions were implemented with a reasonable level of fidelity, then this would imply the reason for the lack of effectiveness was the result of an inadequacy in the underlying theory of change. However, this basic insight would not be supplemented with knowledge of the nature of such failure without studying mechanisms directly. In other words, we would come to the conclusion that the problem lay with theory through a process of eliminating other explanations for the results (e.g. implementation failure, faulty trial design, inappropriate statistical analyses, etc.).

Direct knowledge of why an intervention theory appears to be at fault is important. Such knowledge enables researchers to put forward suggested improvements

to interventions. For example, in some cases interventions maybe unrealistic in that the resources available and the intensity of the intervention may be insufficient to trigger proposed mechanisms, suggesting a greater intensity in some input is required. In other cases, the nature of the intervention may be completely inconsistent with the proposed mechanism. The broader point is that both insufficient intensity or inconsistency could occur in circumstances where an intervention was apparently implemented in accordance with developer intentions. In the case of the trials discussed here, the available evidence suggests implementation was consistent with what might be expected in an effectiveness trial setting. Knowledge of whether the underlying programme theory was at fault, and more specifically the nature of that failure, would have enabled researchers to suggest how the intervention might have been modified or improved through identifying the specific causes of theory failure. One can easily see how such insights would have led to more useful findings for policymakers and schools than simply learning that the interventions were ineffective even though seemingly implemented as specified.

Directing attention toward the importance and study of context

The literature on mechanisms in the programme evaluation tradition is in many cases at pains to stress the contingent nature of mechanisms. The mechanisms that are hypothesised to underpin a particularly theory of change are part of a wider set of factors that act together to produce an effect. Authors such as Cartwright and Hardie (2012) refer to these as supporting factors.

Results from RCTs reveal the consequence of introducing an intervention with its underlying causal mechanisms into a pre-existing context in which the intervention needs to operate in combination with existing factors to produce an impact. If no effect is forthcoming, this could be the result of the absence of key supporting factors which

are presumed in error to be present, or due to the presence of factors that inhibit causal mechanisms from operating. Though our emphasis is on the interplay between context or supporting factors and mechanisms, clearly this discussion has implications for the study of implementation. It leads us to be wary of studies that take no account of the need to modify interventions with regard to the setting or context into which they are introduced.

.

Interestingly, the Paired Reading process study found that the intervention appeared to work less well in settings with a higher prevalence of lower ability children. This type of insight is suggestive of the kinds of findings that might be gained through the systematic consideration of mechanism and their interaction with context. Such a finding hints at an explanation that a key causal mechanism within the intervention was sensitive to the higher prevalence of lower ability pupils in certain school settings. As there was no attention paid to causal mechanisms in the evaluation of Paired Reading, the precise way in which the mix of abilities within a school affected the intervention is unknown. However, we might speculate that one useful insight that might have been forthcoming would be that cross-age peer tutoring interventions need to be adjusted in certain ways in order to work in contexts in which there are a greater number of lower ability pupils. If the causal mechanisms associated with the Paired Reading intervention had been articulated then the particular interaction between causal processes and this aspect of context might have been clearly identified. This would have enabled researchers to be specific in suggesting how the intervention might be altered to make it more effective in such settings.

Contributing to more effective accumulation of knowledge

Finally, we suggest that the study of mechanisms in combination with that of implementation fidelity would have ensured that the results from the trials discussed made a greater contribution to development of future peer education programmes. We argue that placing mechanisms at the centre of trials and thereby giving greater prominence to theory facilitates the accumulation of knowledge over time, and the development and improvement of interventions as further knowledge is accreted.

For example, suppose both elements of implementation and theory had failed in the case of the two trials discussed but that only implementation fidelity has been examined. This would lead us to conclude that implementation failure was the cause of the null findings and we would be in danger of missing fundamental weaknesses in the underlying theory. This would encourage further trials seeking to rectify failings in implementation fidelity that would, even if these weaknesses were addressed, be highly likely to reveal a lack of effectiveness. If both theory and implementation had been studied, failings in both could be addressed in further studies and most importantly progress made. No doubt that where there are aspects of both implementation and theory failure present, it is more challenging to disentangle the relative contribution of these different factors. Nonetheless, studying mechanisms directly is likely to suggest whether the underlying programme theory is viable and plausible in situations in which implementation fidelity is achieved. Moreover, the development of programme theory and the exploration of mechanisms would also produce insights that would be helpful in the development of substantive pedagogical theory, through shedding light on the degree to which processes that are presumed to give rise to effects work or otherwise in real world circumstances.

Concluding remarks

The argument presented in the paper stresses the importance of researchers paying attention to causal mechanisms when evaluating the impact of interventions using RCTs. Clearly the study of mechanisms and indeed implementation present the researcher with a number of challenges, both in terms of theory as well as observation and measurement. We argue that the cause of evidence informed education can only be progressed if these challenges are addressed; and that such difficulties should not deter researchers from grappling with these issues

Interventions implemented within school settings are invariably complex. This complexity limits what can be learnt through RCTs. Trials themselves provide evidence of whether a causal relationship is present but without addition research elements provide only indirect and partial evidence as to how such causal effects are generated. Examining implementation fidelity is necessary but not sufficient for causal explanation. Without directly studying the operation of hypothesised causal mechanisms researchers are restricted in the explanatory accounts they can provide. The degree to which study results can contribute to the on-going development of interventions to support schools in making decisions about which teaching approaches are likely to benefit their pupils will also be limited.

Acknowledgements

The authors acknowledge the funding provided by the Education Endowment Foundation for the case study RCTs discussed in this paper.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173-1182.

Bonell, C., Fletcher, A., Morton, M., Lorenc, T., & Moore, L. (2012). Realist randomised controlled trials: A new approach to evaluating complex public health interventions. *Social Science & Medicine, 75*, 2299-2306.

Bonell, C., Fletcher, A., Morton, M., Lorenc, T., & Moore, L. (2013). Methods don't make assumptions, researchers do: A response to Marchal et al. *Social Science & Medicine, 94*, 81-82.

Britz, M. W. (1989). The effects of peer tutoring on mathematics performance: A recent review. *B. C. Journal of Special Education, 13*(1), 17-33.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford: Oxford University Press.

Collier, D., Brady, H. E., & Seawright, J. (2010a). Introduction to the second edition: a sea change in political methodology. In H. E. Brady & D. Collier (Eds.), *Rethinking social inquiry: Diverse tools, shared standards* (2nd ed.). Lanham, MD: Rowman and Littlefield Publishers.

Collier, D., Brady, H. E., & Seawright, J. (2010b). Sources of leverage in causal inference: Toward an alternative view of methodology. In H. E. Brady & D. Collier (Eds.), *Rethinking social inquiry: Diverse tools, shared standards* (2nd ed., pp. 161-199). Lanham, MD: Rowman and Littlefield.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. (2008). Developing and evaluating complex interventions: the new Medical Research Council guidance. *British Medical Jounral 2008; 337: a1655*.

Dalkin, S. M., Greenhalgh, J., Jones, D., Cunningham, B., & Lhussier, M. (2015). What's in a mechanism? Development of a key concept in realist evaluation. *Implementation Science, 10*(49). doi:10.1186/s13012-015-0237-x

Dunning, T. (2008). Natural and field experiments: The role of qualitative methods. *Qualitative & Multi-Method Research, 6*(2), 17-23.

Education Endowment Foundation. (2015). *EEF evaluation: A cumulative approach*. London: Educational Endowment Foundation.

Funnell, S. C., & Rogers, P. J. (2011). *Purposeful program theory: Effective use of theories of change and logic models*. San Francisco, CA: Jossey-Bass.

Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., & Ngeow, E. (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review, 31*, 79-88. doi:10.1016/j.cpr.2010.09.007

Gerring, J. (2010). Causal mechanisms: Yes, but…. *Comparative Political Studies, 43*(11), 1499-1526. doi:10.1177/0010414010376911

Goldthorpe, J. H. (2001). Causation, statistics and sociology. *European Sociological Review, 17*(1), 1-20.

Hedstrom, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology, 36*, 49-67.

Higgins, S., Katsipataki, M., Kokotsaki, D., Coe, R., Major, L.-E., & Coleman, R. (2013). *Sutton Trust and Education Endowment Foundation: Teaching and learning toolkit: Technical appendicies*. London: Education Endowment Foundation.

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016a). *Implementation and process evaluation (IPE) for interventions in educational settings: A synthesis of the literature*. London: Education Endowment Foundation.

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016b). *Implementation and process evaluation (IPE) for interventions in educational settings: An introductory handbook*. Education Endowment Foundation. London.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods, 15*(4), 309-334.
Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review, 105*(4), 765-789.
Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(1), 5-51. doi:10.1111/j.1467-985X.2012.01032.x
Jamal, F., Fletcher, A., Shackleton, N., Elbourne, D., Viner, R., & Bonell, C. (2015). The three stages of building and testing midlevel theories in a realist RCT: A theoretical and methodological case-example. *Trials, 16*, 466. doi:10.1186/s13063-015-0980-y
Keele, L., Tingley, D., & Yamamoto, T. (2015). Identifying mechanisms behind poicy interventions via causal mediation analysis. *Journal of Policy Analysis and Management, 34*(4), 937-963.
Lendrum, A., & Humphrey, N. (2012). The importance of studying the implementation of interventions in school settings. *Oxford Review of Education, 38*(5), 635-652.
Lloyd, C., Edovald, T., Kiss, Z., Morris, S. P., Skipp, A., & Ahmed, H. (2015). *Paired Reading evaluation report and executive summary*. London: Education Endowment Foundation.
Lloyd, C., Morris, S., Edovald, T., Skipp, A., Kiss, Z., & Haywood, S. (2015). *Durham Shared Maths project. Evaluation report and executive summary*. London: Education Endowment Foundation.
Ludwig, J., Kling, J. R., & Mullainathan, S. (2011). Mechanism experiments and policy evaluations. *Journal of Economic Perspectives, 25*(3), 17-38.
Mahoney, J. (2010). Review Artilces: After KKV: The new methodology of qualitative research. *World Politics, 62*(1), 120-147.
Marchal, B., Westhorp, G., Wong, G., Belle, S. V., Greenhalgh, T., Kegels, G., & Pawson, R. (2013). Realist RCTs of complex interventions - An oxymoron. *Social Science & Medicine, 94*, 124-128.
Maxwell, J. A. (2012) The importance of qualitative research for causal explanation in education, *Qualitative Inqui*ry, 18(8), 655-661
Merrell, C., & Tymms, P. (2005). *InCAS (Interactive Computerised Assessment System): Using individual diagnostic profiles in assessment for learning*. Paper presented at the EARLI Conference, Nicosia, Cyprus
Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. New York: Cambridge University Press.
Oakley, A., Strange, V., Bonell, C., Allen, E., & Stephenson, J. (2008). Process evaluation in randomised controlled trials of complex interventions. *British Medical Journal, 332*, 413-415.
Paluck, E. L. (2010). The promising integration of qualitative methods and field experiments. *The Annals of the American Academy of Political and Social Science, 628*, 59-71.
Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London: Sage Publications.
Pennucci, A., & Lemon, M. (2014). *Updated inventory of evidence- and research-based practices: Washington's K-12 Learning Assistance Program*. Olympia: WA: Washington State Institute for Public Policy.
Peterson, A. (2016). Getting 'What Works' working: building blocks for the integration of experimental and improvement science. *International Journal of Research & Method in Education, 39*(3), 299-313. doi:10.1080/1743727X.2016.1170114

Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal, 45*(1), 206-230.
Robinson, D., Schofield, J. W., & Steers-Wentzell, K. L. (2005). Peer and cross-age tutoring in math: Outcomes and their design implications. *Educational Psychology Review, 17*(4), 327-362.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* New York, NY: Houghton Mifflin and Company.
Tymms, P., Merrell, C., Thurston, A., Andor, J., Topping, K., & Miller, D. (2011). Improving attainment across a whole school district: School reform through peer tutoring in a randomised control trial. *School Effectiveness and School Improvement, 22*(3), 265-289.
Weiss, C. (1997). Theory-based evaluation: Past, present, and future. *New Directions for Evaluation, 76*, 41-55.

Table 1: Results from the Paired Reading and Durham Shared Maths trials – effect sizes on primary outcomes

Paired Reading			Durham Shared Maths		
	Effect size (95% confidence interval)	***Sample size as analysed number of pupils (number of classes)***		***Effect size (95% confidence interval)***	***Sample size as analysed number of pupils (number of schools)***
Overall reading score*			***General maths score§***		
Year 7	-0.02 (-0.15 to 0.11)	1,300 (60)	Year 3	0.01 (-0.07 to 0.09)	2,709 (79)
Year 9	-0.06 (-0.14 to 0.02)	1,265 (60)	Year 5	0.02 (-0.06 to 0.10)	2,598 (79)

Notes:
This table presents results adapted from Tables 6 and 7 in Lloyd, Morris, et al. (2015) and Tables 12 and 13 in Lloyd, Edovald, et al. (2015).
The results presented are adjusted analysis on the primary outcome in both trials and were obtained from multi-level regression models with the primary outcomes as dependent variables.
*Results obtained from a multi-level regression model with random intercepts. Level one in the model is the pupil, level two the class with level three the school modelled as a fixed effects. Covariates included in the model are baseline measure on the outcome (pre-test), student's month of birth, sex and eligibility for Free School Meals.
§ Results obtained from a multi-level regression with random intercepts. Level one in the model is the pupil, level two the school. Covariates included in the adjusted analysis are student pre-test scores, month of birth, sex, ethnicity, English as an additional language, eligibility for Free School Meals and area.