Received: 18 March 2023 | Accepted: 13 December 2023

DOI: 10.1002/berj.3960

ORIGINAL ARTICLE

BERJ BERA

Learning from failure: A context-informed perspective on RCTs

Mike Coldwell | **Nick Moore**

Sheffield Institute of Education, Sheffield, UK

Correspondence
Nick Moore, Sheffield Institute of Education, Sheffield Hallam University, Sheffield S1 1WB, UK.
Email: n.moore@shu.ac.uk

Funding information
Education Endowment Foundation; Bell Foundation; Unbound Philanthropy

Abstract

Discussions of randomised controlled trials (RCTs) in education that do not show an impact regularly focus on the intervention and how it failed to impact on expected measures, with typologies identifying persistent critical points of failure. This paper uses one such RCT—the Integrating English programme—to exemplify the application of a new model to explain failure in RCTs. To do so, the paper develops a set of categories of context drawing on the wider social evaluation field: backdrop, design, operation and interpretation. Thus, the paper exposes critical weak points in the commission and interpretation, as well as the implementation, of an RCT. Our aim is to work towards more robust evaluations by demonstrating that it is not simply the programme design, implementation and evaluation that can contribute to a lack of impact; there can be more fundamental system issues at play.

KEYWORDS

programme development, randomised controlled trials, selected contextual issues

INTRODUCTION

Randomised controlled trials (RCTs) are becoming commonplace in the United Kingdom and internationally: in England alone, at the time of writing, the Education Endowment Foundation (EEF) had funded over 200 trials since 2012 and it is claimed that about half of English schools had been or were currently engaged in some way with EEF RCTs by 2019 (Nevill, 2019), so presumably the current figure will be even higher. However, it is common

Mike Coldwell and Nick Moore are equally contributed to this work.

Key insights

What is the main issue that the paper addresses?

RCTs are central to producing evidence of 'what works' in schools, but most produce 'null' results; mostly nothing works. Explaining this failure can improve educational research. We consider how the contexts of the evaluation may be as important as the programme when interventions fail to impact education.

What are the main insights that the paper provides?

To learn from RCT failures, we consider the backdrop to the intervention, its design, implementation and interpretation. We contend that statistical measures obscure these subjective decisions. We analyse how policy priorities, evaluation preferences and education practices may account for the failure to show impact in our case study.

for educational interventions evaluated by RCTs to show no effect compared with a control group; as discussed in the next section, Lortie-Forgues and Inglis (2019) present evidence that less than a fifth of EEF trials showed evidence of impact. There is a large recent literature critiquing RCTs, from a range of perspectives. Some of these critiques point to philosophical issues with trials, with concerns that they cannot represent real practice, or focus on only some kinds of change processes, or that their assumptions about causation are simplistic (see, e.g., Pawson, 2013; Wrigley, 2018). Others—as we go on to discuss in this paper—point to technical issues related to implementation and design (e.g., Lortie-Forgues & Inglis, 2019). Such criticisms draw on the theory, method or implementation of the intervention as explanations for their failure to show impact (see Burnett & Coldwell, 2021 for an extended recent discussion).

This paper draws on the critiques outlined in the previous paragraph, but we come from a different direction. The paper develops and utilises a frame derived from social evaluation literature on forms of failure and contexts for evaluation to consider not just how an intervention can show no impact or 'fail' in its evaluation, but rather how the system within which an RCT takes place can fail the intervention itself. We then apply this frame to focus on one programme which, under an RCT evaluation, did not demonstrate impact. Thus, we address the key question for the paper: how can taking a context-informed perspective on 'programme failure' improve learning from RCTs?

In the remainder of the paper, we set out some fundamental features of RCTs, focusing on the understanding of causation in such trials, before introducing and developing the theoretical tools used to explore how a programme can 'fail', and provide contextual explorations of this failure. In so doing, we present a new model that exposes potential points of failure in the process of evaluating educational interventions with an RCT. We then briefly describe the case to which we apply this frame, the Integrating English programme, its trial and key findings. We then analyse the programme to test the validity of the proposed model. We conclude by highlighting key issues for situating future evaluations, and RCTs in particular, in their broader social context, and warn of critical points where evaluation design can fail interventions.

Before moving on, it is important to be clear about our use of the term 'failure'. The result of an RCT can show a positive, null or negative effect. When negative it may—in the RCT field—be called a failure. A null result is not usually thought of as a failure in this field, and can be seen positively, as a demonstration that the programme under evaluation

is no better than whatever else could be done as an alternative. However, in the social evaluation literature, both negative and null results are often referred to as representing failure (Coldwell & Maxwell, 2018; Stame, 2010), that is a failure to show impact. This is the meaning of failure used in the paper. As we go on to discuss, we propose to extend the idea of failure beyond the trial and programme, considering the wider context within which RCTs play out.

UNDERSTANDING CAUSATION AND FAILURE IN RANDOMISED CONTROLLED TRIALS

In the United Kingdom, RCTs gained popularity following Goldacre's series of newspaper articles, culminating in the Haynes et al. (2012) report to the Cabinet Office. The United Kingdom's What Works Centre for Education, the EEF, was set up in 2011 and reflects the work of the National Center for Educational Evaluation and Regional Assistance (NCEE) in the United States. The EEF has put considerable investment into education research, undertaking about 200 RCTs at the time of writing, within which (from 2011 to 2019) 'over 12,000 schools, nurseries and colleges and 1.2 million children and young people' were engaged (EEF, 2019). Similar What Works Centres in other policy areas promote evidence-informed policymaking: 'What Works is based on the principle that good decision-making should be informed by the best available evidence. If evidence is not available, decision-makers should use high quality methods to find out what works' (What Works Network, 2019). The preferred method of finding the best evidence for these centres, to date, has been RCTs. In this section we begin by focusing on the logic of causation implied in RCTs, and highlight the shifting perspectives on the scientific agenda and understanding of causation implied within them. This then leads to a set of explanations of failure in such studies.

Causation, RCTs and the EEF evaluation model

RCTs depend, often implicitly, on a predictable causal relationship. The statistical models of the majority of RCTs identify whether there is a non-random relationship between a change and a specific variable. In the case of EEF trials like Integrating English, a change is typically introduced into schools and a variable, such as scores in a standardised assessment, is compared to schools with similar characteristics that did not experience the same change. If the variable can be measured credibly (Rogers, 2011), and if the comparable group shares all important features, the logical inference is that the change is the factor that caused the difference to the variable. Consequently, assuming sufficient statistical power to generalise these findings, the change can be considered to be 'what works' in similar contexts, if in such contexts the same change can be implemented.

This robust model offers strong evidence for causation. The key weakness of an RCT is that it only demonstrates *if* an impact occurs; it is unable to link the change in practice to how or why the variable experienced an impact. In place of a realist paradigm to discover 'what works for whom in what circumstances' (Pawson, 2013, p. 15), the EEF retain a positivist approach through an implementation and process evaluation (IPE) intended to measure the key variables in a theory of change (Humphrey et al., 2016). Initially conceived as a means to uncover the conditions within which an intervention is likely to gain the greatest likelihood of success and encourage fidelity to these, the IPE has become a key aspect of the evaluation process for the EEF, with increasing emphasis on contextual variation and theory-based evaluation (Pawson & Tilley, 1997; Rogers & Weiss, 2007). We discuss the IPE throughout this paper.

Under the EEF's scheme, evidence of impact progresses a programme of change from pilot, through efficacy and effectiveness trials, to scale-up and advocacy (EEF, 2023a). However, if at any stage the programme evaluation does not provide evidence of a positive impact, it is likely to be 'delisted' and no longer supported by the EEF; the aim of the programme designers to provide evidence of the efficacy of the programme will not be achieved.

Here we begin to link the RCT conceptualisation to the wider evaluation field, which uses a more brutal terminology: with no evidence of impact, the programme will have failed its evaluation (e.g., Stame, 2010). In fact, RCTs typically do not show a significant impact, particularly those funded by the EEF (Lortie-Forgues & Inglis, 2019). To be more precise, Lortie-Forgues and Inglis found that only 18% of 119 EEF trials in the United Kingdom and 29% of NCEE trials in the United States showed a positively significant effect (23% overall). In an alternative Bayesian analysis, 20% of EEF trials and 27% of NCEE trials (23% overall) produced results that support the alternative null hypothesis. That is, in about 80% of EEF trials, the data was either uninformative or showed no impact (Lortie-Forgues & Inglis, 2019). This is of major importance for the EEF, and the use of the public finances that support them.

These 'null results' (Jacob et al., 2019) are visible because of the EEF's and NCEE's laudable policy to publish all results, but how do we explain this high level of failure to demonstrate impact? We first turn to the wider evaluation literature to develop the explanation.

Failure in evaluation

According to Stame (2010), an intervention may fail its evaluation in one of four ways (see Table 1). First, the intervention may not be implemented 'correctly', as prescribed by the programme designers, leading to **implementation failure**. Here, evaluations can measure the amount of 'fidelity' to the programme's goals and methods to enact change. Implementation failure occurs because the intervention is not implemented in a way that is likely to produce the desired effect.

If the evaluation measures a change in the 'causal variable'—the expected change—but the theory cannot explain that change, the programme experiences a **theory failure**, because it cannot make predictions about change under new conditions. A programme theory provides 'a description of how an intervention leads to change' (Coldwell & Maxwell, 2018, p. 269) and fails when it is unable to do so.

Programme failure occurs because the intervention does not achieve its intended impact—the change in the causal variable does not occur. In rejecting theory failure as the reason for an intervention failing to show an impact, the evaluators must consider whether

TABLE 1 Types of evaluation failure.

Type of failure	Characteristic
Implementation failure	The programme is not implemented as directed and/or uniformly
Theory failure	The theory behind the programme fails to explain the connection between the 'causal variable' and the programme
Programme failure	The programme (e.g., extra homework) fails to have an impact on the 'causal variable' (e.g., GCSE maths scores)
Method failure	The research design fails to find the link between the programme and the 'causal variable'

the programme may succeed using a different design. If so, this is considered a programme failure (Nielsen et al., 2006) and the IPE is often a good place to identify where there may have been inadequacies in the programme under evaluation (Rogers, 2014). In such cases, a revised trial is the appropriate response, making changes to aspects of the programme to better conserve the causal relationships in the theory of change.

When the evaluation design is unable to identify the causal link between the intervention and the anticipated change, this may be due to **method failure**. Within a cause–effect paradigm, a method failure may occur because it is unable to correctly reject a null hypothesis (Stame, 2010) (a Type II error) or falsely rejects the null hypothesis (a Type I error).

Stame's forms of failure have proved valuable in many contexts; however, they have only very limited ability to fully *explain* how these failures occur. This is the focus of the next section.

FRAMING THE EXAMINATION OF FAILURE: DEVELOPING CONTEXT CATEGORISATION

In this section, we subject the context to much greater scrutiny, drawing on and extending Rog's (2012) framing and then linking these context categories to provide an explanatory framework for understanding Stame's forms of failure. The aim of this is to broaden our understanding of context beyond the immediate environment within which a programme or intervention takes place—the usual understanding of context—to encompass the framing of the problem, the nature of the intervention, the wider funder's goals and evaluation choices, and more. Rog (2012) delineates five categories of context, as follows.

- The **problem context**: the nature and framing of the problem or phenomenon under study.
- The **decision-making context**: the funder's motivations, requirements and understandings.
- The **intervention context**: the structure, maturity and complexity of the intervention.
- The **evaluation context**: the constraints on the evaluation design (time, budget, required methods).
- The **broader environment/setting**: the setting for the intervention—the usual meaning of context in evaluation contexts.

Rog (2012) recognises that aspects of these categories of context may overlap and have mutual causes and effects. However, we would add that some aspects of the context of an evaluation are still missing from this categorisation, based on recent reviews (Coldwell, 2019; Greenhalgh & Manzano, 2021; Nielsen et al., 2021). In this paper we pick out three additional categories.

Firstly, since Rog was writing, the field of implementation science (IS) has developed. Initially derived from the health field, IS was first conceptualised as 'the scientific study of methods to promote the systematic uptake of research findings and other evidence-based practice into routine practice and, hence, to improve the quality and effectiveness of health services' (Eccles & Mittman, 2006, p. 1). It has since been extended across policy fields. Therefore, the implementation context is important—how the evaluation is implemented, or realised, in practice in the settings (e.g., schools) engaged in an RCT. We call this phase of the evaluation the **operation** and divide it into two parts: the **teaching** context and the **broader environment**. It is in the teaching context that evaluations must identify how models of training may change practice (Boylan et al., 2018; Opfer & Pedder, 2011; Strom et al., 2023), the effects of balancing fidelity with adaptability

(Lendrum & Humphrey, 2012) and the impact of school-based policies and practices. These factors are likely to impact all interventions and will also be influenced by national testing and inspection regimes, for example. The national and international educational environment constitute the broader environment in which interventions may or may not succeed.

Secondly, the theory of change that is more or less implicit (see Coldwell & Maxwell, 2018) in any intervention also determines how an intervention will be designed, creating a **causal theory context** that will have a significant but varying degree of influence over an evaluation (which relates directly to Stame's 'theory failure').

Thirdly, we would add that the consequences of an RCT impact study are not confined to the study itself, and that the use of the results of an RCT for a What Works Centre are part of the array of contextual factors that influence the design and implementation of a trial. To put it another way, a **consequential context**, including the *interpretation* of success or failure of a study, must be considered as part of the design of an evaluation, just as consequential validity needs to be central to the design of a high-stakes test (Messick, 1989). RCTs and evaluations do not exist in a vacuum, but influence the theory, practice and policy environments (Burnett & Coldwell, 2021) and contribute to the construction of social reality (Law & Ruppert, 2013).

In Table 2 we map Stame's (2010) four categories of failure against these eight contextual factors, drawing them together into a set of four overarching 'context categories' which map roughly onto the phases of an evaluation, as a *context-informed explanatory framework*, to provide a structure for researchers to investigate the reasons behind the success or failure of an RCT. While the framework allows us to identify individual features, these features often influence each other.

Backdrop

We organise the context categories in Table 2 to broadly reflect the timeline of an intervention and evaluation. We start with problem and decision-making contexts, as these frame the inception of the intervention. They are the 'backdrop' in that they 'pre-exist' any intervention or evaluation. An intervention project typically attempts to address a problem that has previously been recognised and prioritised by policymakers and funders, and this forms a key reason for funding the project. Priorities may be influenced by political agendas, charitable body objectives and charters, and the concerns of research and media bodies.

TABLE 2 The context-informed explanatory framework: mapping evaluation design and context to forms of failure.

Context categories/ phases	Contextual factors (developed from Rog, 2012)	Forms of failure (Stame, 2010)			
		Theory	Programme	Method	Implementation
Backdrop	Problem				
	Decision-making				
Design	Causal theory				
	Intervention				
	Evaluation				
Operation	Teaching				
	Broader environment				
Interpretation	Consequential				

The problem context will be framed by the political, social and economic priorities in the time and place in question. In the United Kingdom, for the past 20 years there has been an increasing focus on English and Mathematics education, for example. The EEF frame all of their trials as being concerned with narrowing achievement gaps between more and less advantaged pupils. Thus, improving English and Mathematics outcomes is a regular feature of EEF trials, as the problems they are seen to try to solve are framed in this way. It is important to note that the problem context can contain many unquestioned assumptions. In the United Kingdom, for instance, there is currently little public debate around the division of schooling into primary, secondary and tertiary, or into government and private funding. Politically motivated priorities, trending topics of research and aspects of education amenable to 'interventionisation' (Burnett & Coldwell, 2021) are likely to be funded before studies that may provide significant challenges to power structures, topics with low public interest and educational innovations that are highly context-specific or have outcomes that may be difficult to measure. The funding of any RCT, and any measure of impact, should be compared to the opportunity cost of not funding other projects or investigating educational change through other means. The backdrop also includes the impact from prior studies on the problem context, as this informs the policy environment, creating a dynamic cycle of evolution and change.

The decision-making context is also a determining factor in a trial. Funders typically have preferred ways of working, qualifying criteria for projects and operational processes that have proved successful in their experience, and most demand particular methodological approaches. For example, the EEF (2022a) have completed more projects (73) in Key Stage 2 (KS2; 7 to 11-year-olds) than any other stage, and combined with Key Stage 3 (11 to 14-year-olds), these EEF projects outnumber Early Years, Key Stage 1, Key Stage 4 and Key Stage 5 by almost two to one (122:71). While most funders will have a *modus operandi*, they may also be open to challenge and to considering alternative approaches when the case is made (which is likely to be mitigated by the status of the trial developers and evaluators). Thus, the decision-making context is not entirely predictable, but its negotiation will have an impact on both the intervention and the evaluation design.

Design

An intervention and its evaluation are purposefully designed by considering the implicit or explicit theory of change in the project, the context of the intervention and the available options for evaluation. This design phase forms a critical part in the evaluation of any trial and is often the focus of considerable negotiation between funders, project leaders and evaluators. In the causal theory context, the negotiation of a theory of change for evaluation (Breuer et al., 2015; Coldwell & Maxwell, 2018; Connell & Kubisch, 1998) will determine: the key elements of an intervention that provide the frame for implementation with fidelity, for the purpose of replication and expansion; the causal mechanism(s) that need to be monitored by a process evaluation; the contextual factors considered influential; and what results are valued and how they may be measured for impact.

As well as influencing a funder's choice of type of trial (e.g., pilot, efficacy, effectiveness trail; EEF, 2023a), the intervention context will also vary depending on how well established the concepts, practices and theories are in the broader context (in the backdrop).

The evaluation context changes the methodological and practical decisions negotiated during the design of the evaluation through factors including budget and access to resources, and the intersection of experience, preferences, beliefs and ideologies in the different parties. Therefore, an evaluation is most likely to be a compromise not just of sampling but of data type and research design.

Operation

Having designed an intervention and an evaluation, the programme is operationalised. The real-world teaching context may differ considerably from the design, and is therefore likely to be adapted to the schools, classrooms, teachers and students that experience the intervention. One key objective of an RCT is to report on a trial's ability to impact the real-world environment, and the IPE has become a key tool in identifying which aspects of these real-world contexts are critical to successful adoption and integration of a change programme (Humphrey et al., 2016).

Although it is almost impossible to draw a line between what happens inside an institution such as a secondary school and the policy, regulatory, social, economic and theoretical forces that shape it, each school exists within and contributes to this broader environment, responds and adapts to it, and is in part its product. The context of the broader environment may apply equally to schools within the same region, but their internal characteristics will produce an individual response that evolves with the internal and external dynamics of these interacting complex systems (Jacobson et al., 2019; Maxwell et al., 2022).

Interpretation

For most evaluators, the publication of a final report is the end of their contract with the funder and may signal the end of contact with the programme designers. However, the report and its findings begin their life and their impact at this point and contribute to the consequential context. The publication of findings will take place within a context of previous research, current policy and political priorities. The context cannot change the results of the trial, but the way that evaluators and funders present results will respond to the context. Therefore, at this point the consequential context forms part of the backdrop context for subsequent studies, creating a dynamic cycle of evolution and change.

THE CASE STUDY PROGRAMME AND ITS FINDINGS

In this section we introduce an RCT that will be used as a case study for applying the context-informed explanatory framework described above. Integrating English was developed from the Language in Learning Across the Curriculum (LiLAC) programme (Custance et al., 2012) and adapted from its original Australian context for the United Kingdom by developers based at Enfield Council. It utilises a functional approach to linguistics and grammar, aiming to break down the process of teaching the language of a curriculum subject to all pupils, focusing on the features of language and meanings that are made in the genres that constitute each school subject. Culliney et al. (2019, p. 8) give examples of what this may mean in different subjects: 'teachers may focus on the generic and grammatical features of verbal art and everyday rhetoric (English); classification, experimentation, and reporting (science); recounts and causation (history); or problems, explanations, and proofs (mathematics), and so on'. The programme team from Enfield Council added United Kingdom-specific content to the training programme, focusing particularly on supporting pupils with English as an Additional Language (EAL), including classroom 'hints and tips' and specific techniques. An important element of the programme was the development of a scheme of work, produced by teachers as part of the latter stages of the training and then further developed back in school to structure classroom teaching.

The trial: Integrating English

Integrating English was the first RCT evaluation of a programme that uses a systemic functional linguistics (SFL) and genre pedagogy approach in the United Kingdom. SFL holds that education, particularly within the disciplines, is largely a process of learning language, learning through language and learning about language (Halliday, 1993). All students are expected to improve with genre pedagogy, but those regularly disadvantaged by the dominant education system can narrow the attainment gap by being exposed to how language construes and enacts the genres of the disciplines (Rose & Martin, 2012). Details of the programme and the evaluation methodology can be found in the evaluation report (Culliney et al., 2019).

The programme ran over two school years, involving training for teachers of Year 5 children (aged 9–10) who then taught the class for a term, with further training for teachers of the children when they moved into Year 6 (aged 10–11), with further teaching taking place at that point (see Table 3).

The programme was evaluated using a two-arm, school-level clustered RCT, with school-level randomisation. In simple terms, this meant that a number of schools were invited to take part in the trial and then randomly allocated either to the intervention group or to a control group, which did not take part in the LiLAC programme (although they received £200 as a thank you for engaging in the evaluation elements). Because of the focus on EAL, schools were invited to take part in the project if they had a minimum of ten EAL pupils and at least two classes in Year 5. A total of 91 schools registered to take part, and after an initial baseline test of all pupils in Year 5 across the sample (the GL Assessment Progress Test in English (PTE) level 9, an age-appropriate, standardised assessment of technical (spelling, grammar and punctuation) English skills), 45 schools were randomly allocated to a control group and 46 to the intervention group. After dropouts, 1817 pupils across 39 intervention group schools and 1790 pupils across 41 control group schools were included in the trial. Pupils across all schools were tested after the intervention period using a KS2 writing paper as the primary outcome measure, with KS2 writing and grammar scores as secondary measures and the test results compared.

The trial was accompanied by an IPE designed to understand how the programme was experienced in schools. The IPE included an observation of six training events, visits to a sample of 14 schools from five regions to observe classes and interview trial teachers and school leaders, and collection of documentary analysis of action plans and schemes of

TABLE 3 Programme schedule.

	Year 5 teachers	Year 6 teachers	Pupils (2016/2017 Y5 cohort, Y6 in 2017/2018)
Spring term 2016/2017	1. Receive LiLAC training 2. Write scheme of work		
Summer term 2016/2017	3. Implement scheme of work	4. Receive some LiLAC training	Teaching experienced by pupils in Y5
Autumn term 2017/2018		4. (continued) receive LiLAC training 5. Write scheme of work	
Spring term 2017/2018		6. Implement scheme of work	Teaching experienced by pupils now in Y6

Note: Reproduced from Culliney et al. (2019, p. 11).

work that were requested from all schools in the intervention. Further details are provided in Culliney et al. (2019).

The findings

Starting with the RCT design as the main impact evaluation, analysis of the primary outcome provided no evidence that Integrating English improved pupils' KS2 writing outcomes. The sample size, fidelity and dropout rate meant that—using the EEF's rating system—this result was judged to have a moderate to high security rating. Further analysis was undertaken of subgroups, examining differences in KS2 writing outcomes for pupils receiving Free School Meals (FSM) in the control and project groups (used as a proxy for social disadvantage), and comparing writing outcomes for EAL pupils (who were expected to see a more positive outcome). In each case there was no evidence of impact, although lower security ratings were assigned due to the smaller numbers of pupils involved. These results are presented in summary form in Table 4, and explained in more depth in Culliney et al. (2019, pp. 25–29). In short, according to the RCT the programme failed to show an impact.

Selected results from the IPE are provided here to explain or contextualise the RCT results (Humphrey et al., 2019) and to frame our discussion of the failure of the Integrating English trial to show impact. Overall, interview evidence showed that teachers responded very positively to the training programme, and many provided examples of 'cascading' professional learning to colleagues. Observations and interviews revealed a wide range of positive effects on pupil performance (e.g., quality of writing across the curriculum), as well as teacher development, particularly in relation to understanding and tailoring teaching to EAL learners. A key finding was that many teachers were grateful to have techniques and activities that would better serve their EAL pupils, as most claimed their previous training had not prepared them adequately for this challenge. For some teachers, this was the main learning from the programme.

There are also findings in the IPE to help explain the findings from the RCT. For instance, fidelity to the programme was modest. Less than one-third of all trial schools completed all three indicators of fidelity: high attendance at training; a school-level action plan to allocate resources to the programme; and related schemes of work for Years 5 and 6. The indication from the interviews was that close to 100% of teachers were keen to talk about genre pedagogy—the fundamental concept behind the programme—but the other three key concepts were less well known, and evidence from observations and from the analysis of schemes of work also revealed that the understanding of these concepts rarely translated into classroom practice or planning. That is, it was uncertain whether the enthusiasm for the programme translated into real development in teacher cognition and practice and, therefore, according to the theory of change, whether there would also be a significant impact on pupil performance.

In summary, the RCT design failed to show an impact on EAL students' general writing ability, on economically disadvantaged pupils and overall and so, despite the highly positive reaction from teachers, the intervention was a failure. The following section uses the context-informed explanatory framework (see Table 2) to investigate where the intervention may have failed the evaluation and where the evaluation may have failed the intervention.

CONTEXTS OF INTEGRATING ENGLISH: LEARNING FROM FAILURE

This section applies the analytical model to the experience of the evaluation team of the Integrating English intervention. We examine our role as evaluators in the RCT for Integrating

TABLE 4 Summary results of the Integrating English impact evaluation.

	Raw means				Effect size		
	Intervention group		Control group				
Outcome	*N* (missing)	Mean (95% CI)	*N* (missing)	Mean (95% CI)	*N* in model (intervention; control)	Hedges *g* (95% CI)	*p*-Value
Overall	1817 (636)	15.72 (15.45, 15.99)	1790 (489)	15.79 (15.5, 16.08)	3607 (1817; 1790)	−0.05 (−0.21, +0.12)	0.577
EAL subgroup	995 (298)	15.87 (15.51, 16.23)	983 (274)	15.99 (15.6, 16.37)	1978 (995; 983)	−0.06 (−0.25, +0.12)	0.517
FSM subgroup	630 (271)	14.77 (14.32, 15.23)	661 (198)	14.91 (14.43, 15.38)	1291 (630; 661)	+0.01 (−0.18, +0.20)	0.917

Note: Derived from Culliney et al. (2019, pp. 26–28).

English, an intervention managed by a team from Enfield Council and funded by the EEF, the Bell Foundation and Philanthropy Unlimited. We consider how decisions and actions taken at each stage of the evaluation process may have affected the RCT and the ability of the programme to demonstrate an impact. We point out, using the focus of critical and contributory factors, how the evaluation may have failed the programme as much as the programme failed the evaluation. We propose that the purpose of the context-informed explanatory framework is to analyse and identify likely and potential points of evaluation failure (*contributory* factors), highlighting where factors may determine the outcomes of the evaluation (*critical* factors). The analysis is presented as illustrative rather than comprehensive due to the limitations of space.

Backdrop to the evaluation of Integrating English

A review of the full context of education in primary schools in England is beyond the scope of this paper, but here we highlight some of the most relevant factors in the backdrop that relate to the Integrating English trial. Schools in England, where this trial was conducted, face a range of financial and re-regulatory pressures, including OFSTED inspections and 'league tables' for high-stakes exam results (Greany & Higham, 2018; Supovitz, 2009). Combined with the push towards the 'indentured autonomy' of academisation (Thompson et al., 2021) and its corresponding transfer of public money to private hands (Wilkins, 2017), there is an increase in teacher attrition and difficulties in delivering continuing professional development (CPD) to encourage staff retention and growth (Long & Danechi, 2022). In this environment, the offer of cash for engaging in EEF trials and the CPD typically associated with them can solve a range of problems for school leaders and could contribute to the problem context for EEF-funded RCTs.

The Integrating English intervention was commissioned by the EEF because it matched the requirements of a round of funding for projects that focused on pupils in schools in England that use EAL. That funding call was made in the context of a number of key indicators identified by prior research. 'The percentage of pupils in English primary and secondary schools aged 5–16 who are recorded as EAL has more than doubled from 7.6% in 1997 to 16.2% in 2013' according to Strand et al. (2015, p. 5), while the evidence base for what works for these pupils was sparse: of 29 studies that included measures for 'research design, sample size, level of participant attrition, and fidelity and validity' (Murphy & Unthiah, 2015, p. iii), only two were conducted outside the United States, where attitudes to heritage languages and support for EAL in schooling (De Costa & Qin, 2015) are very different from the UK context of Integrating English. Meanwhile, the financial and other support available to schools for these students has declined considerably since 2000 (Hutchinson, 2018), and so there was (and remains) much more to understand about how schools can ensure EAL students in UK schools are not disadvantaged at school because of linguistic competence. Defining the problem context in this way determines the type of programme that could be considered to qualify for this funding, and so is a contributing factor in the potential success of the evaluation.

While closing the attainment gap by focusing on what works for the increasing EAL school population may constitute the problem context, the decision to implement an RCT determines a significant part of the decision-making context and contributes to the possibility of method failure. For instance, to allow for school-level randomisation, schools that did not have at least 10 EAL pupils across two classrooms were excluded from the study. As a What Works Centre, the EEF's mission includes supporting 'teachers and senior leaders to raise attainment and close the disadvantage gap' (EEF, 2022b). Educational gaps disadvantage pupils intersectionally across gender, race and ethnicity, disability and class lines, and

yet no formal system exists to monitor pupil performance by social class in UK schools, as this is not a protected characteristic. Proxy measures, such as FSM or Pupil Premium, are used to approximate indicators of this central discriminator in modern British society. This can only make evidence-informed approaches to closing the attainment gap that depend on segregating the population inherently less valid. While these trials may often produce high-quality results of impact, this paper and other approaches question its ability to respond to all educational and research contexts. Alternative approaches, such as the 'what works for whom in what circumstances' realist evaluation paradigm (Pawson, 2013, p. 15), would create a very different decision-making context for the trial.

Design of the Integrating English RCT

The three context elements of the design phase are identified in Table 2 as **causal theory**, **intervention** and **evaluation**. Turning first to the underpinning causal theory, in the Integrating English evaluation we identified two related causal processes: firstly, a professional development programme that was designed to lead to teacher practice changes; and secondly, teacher practice changes based on using a systemic functional linguistics approach that was intended to lead to positive pupil understanding and use of language. From an evaluation perspective, the evaluation could say little about these causal processes as, firstly, the operation issues identified in the next section meant that the CPD programme was variably conducted and so the theory was not fully tested. Given that there were variable levels of engagement and responses to the CPD element, the practice changes were also variable. In short, that the evaluation was able to say little in relation to the value of the underlying causal theory is a critical theory failure in this evaluation. Culliney et al. (2019, p. 52) summarised this as follows:

> … whilst the prior Australian research evidence summarised in the background section indicates that the underlying theory in relation to the potential usefulness of genre pedagogy approach has a clear theoretical basis and some empirical corroboration, we have much less evidence in relation to this second causal process from the current study. Since we did not see the changes in teachers' practices, we have limited evidenced (sic.) on whether these practices would have led to positive pupil change had they been seen.

The intervention and evaluation contexts, however, both provide more productive lenses to understand reasons for failure. In the previous section, we outlined how decision-making and problem contexts feed into the intervention and evaluation design. The focus of the funder on RCTs was the main feature of the evaluation context, and this led to a series of decisions about the design of the evaluation and the operation of the programme which increased the likelihood of intervention failure. Firstly, there was a shift in emphasis early in the process from an initial intended focus on secondary schools to primary schools. The Bell Foundation was keen to utilise a cross-curricular approach in secondary schools, since the focus of the intervention—the development of subject literacy—takes on more significance in secondary schooling (Fang & Schleppegrell, 2010; Unsworth et al., 2022; Veel, 1997). However, this was problematic as it was likely to require a very large number of staff teaching across the subjects, creating logistical challenges to replace them during training days. This led to a decision to focus on primary schools.

The second area in which the backdrop contexts influenced the intervention design related to the need to sequence the programme around the organisation of the school year and—in particular—the focus in upper primary schools on KS2 tests (known as 'SATs'),

intended as the main measure of the impact of the intervention. To encourage schools to engage in the programme, the sequencing made sure to avoid the final term of Year 6, which had the effect of splitting the programme—which in its original format took place over a single year—over two school years, with operation impacts as discussed in the next section.

Turning back to the evaluation context, RCTs require well-defined outcome measures (see 'Failure in evaluation' section). For this intervention, the primary writing score in the KS2 tests was used as the main outcome measure, and this created a number of issues relating to concerns about its validity as an appropriate measure by which to judge the programme's impact. Interviews in the IPE evaluation with teachers and the programme delivery team questioned the relevance of a writing test for English to a cross-curricular programme; the evaluation report noted 'the variety of curriculum subjects that Integrating English was applied to in different schools would make it next to impossible to find a suitable writing test for all' (Culliney et al., 2019, p. 53). Secondly, the measures were redesigned during operation when the statutory Spelling, Punctuation and Grammar test was made optional, which may have affected the engagement in the test. These issues were drawn together in the evaluation as a concern that an appropriate tool that 'accurately measured responses in a comparable way in relation to the specific subjects in the range of subjects addressed in schools was not available (and it is doubtful such a tool could be created, not, at least, without a huge amount of developmental work). Without such a tool it is difficult to measure literacy development across subjects' (p. 53), which could be a critical factor in method failure in this evaluation.

Finally, the report identified a further problem with using standardised KS2 tests, particularly when using 'coarse' grades as outcome measures (Smith et al., 2021). Given their status as a key element of the strong accountability system in English schools, as indicated above, there is inevitably a huge amount of activity taking place to maximise scores in KS2 tests, particularly in the final term of Year 6, and these are very likely to 'drown out' any possible influence of a single programme which is not directly designed to maximise such scores. Consequently, the impact evaluation shows us that the 'Integrating English approach produced no gain in KS2 GPS and KS2 Reading scores over these other methods' of test preparation (Culliney et al., 2019, p. 53) due to a critical factor within the evaluation context failing to measure adequately for causal change, and contributing to method failure.

Operation of the Integrating English RCT

When the Integrating English trial was implemented in schools in five regions across England, as part of the criteria for fidelity to the programme the evaluation design demanded that: (1) schools produce an action plan to show how the programme would become part of their operations; (2) teachers produce schemes of work identifying when and how they would practice elements of the programme in classrooms at least twice a week; and (3) all teachers would attend a minimum of four out of five training days. While training completion was very high (92%), fidelity to planning Integrating English was low. School-wide action plans were submitted to evaluators by 54% of schools, schemes of work were submitted by 49% and only 31% of trial schools submitted both forms of evidence. It is possible that the submission of documents was too onerous, or that Integrating English was being implemented without documentary evidence, or that in the majority of schools the implementation of Integrating English was not fully supported by school leaders or teachers. However, fidelity measures are likely to be critical factors in the teaching context of an intervention and increase the likelihood of implementation failure. Further evidence is required to determine if this or other factors resulted in a low return of fidelity indicators for most trial schools involved in Integrating English.

A significant feature of the Integrating English programme, the LiLAC approach and genre pedagogy is the use of a teaching and learning cycle following these stages: setting the context, modelling and deconstruction, joint construction and independent construction (Custance et al., 2012; Rose & Martin, 2012). These stages enable awareness-raising of features of different disciplinary genres to take place in a supportive environment before the genre is practised as a group and then individually. In the schemes of work that were submitted and the observations in case study schools, there was no evidence that this cycle was implemented in classrooms. This may be the result of misunderstanding during training, an incompatibility with classroom practice in schools in England, or a lack of understanding or confidence to identify features of genres. Further evidence is required to determine if these or other factors meant that the evidence for full operation in schools was insufficient, but evaluation data points to a mismatch between the operation, including the teaching context, and the programme's aims, representing a critical factor in implementation failure.

Pressures on the evaluation design produced a programme spread over two school years (see 'Design of the evaluation of Integrating English' section). Because of the impact of standardised tests in Year 6, the evaluators, funders and programme team agreed that the programme would be largely obscured by exam preparation in the final term of Year 6. However, it was also recognised that any changes in writing could be lost between Years 5 and 6. Consequently, the programme was implemented over Years 5 and 6, so that the same pupils experienced Integrating English in the last term and a half to two terms of Year 5 and the first two terms of Year 6, often with different teachers. This design created multiple opportunities for increased variation in teaching styles, focus of the intervention, use of different elements of the programme and other differences. The extent of this variation due to the context of the broader environment is almost impossible to measure, creating an even more complicated picture of the programme, but is likely to contribute to method failure.

Interpretation of the Integrating English RCT

In this case, the immediate consequence of providing no evidence of an impact was stark: the summary on the funder website states 'The EEF has no further plans to trial the Integrating English programme' (EEF, 2023b). That is, although there was evidence from the IPE that the programme showed some success, despite low fidelity, the impact of the programme on student performance was the only measure that mattered. Beyond this, it is unclear. Enfield Council, as the delivery partner, ceased to provide the programme; so to all intents and purposes it is over. There are potentially negative effects for the underlying LiLAC programme; certainly, the evaluation outcome will not help attract further funding. Moving further out, as the only RCT that has, to date, tested the principles of systemic functional linguistics and genre pedagogy, the trial result provides little support for further impact evaluations; the same claim might equally be made for interventions focused on EAL. These are potential consequences; but what can be said without doubt is that the results of an RCT which is not replicated are not positive for any of these areas. More broadly, the same could be said of the 72% of trials that produced no evidence of an impact. It is this domain of consequence which is, in one sense, the topic of this paper—and we return to this in the concluding discussion section.

Drawing together

Appendix A presents a summary of our analysis of the Integrating English RCT, showing which factors we identified as contributory or critical factors in the different forms of failure.

We offer the context-informed explanatory framework as a tool to identify potential points of failure in future RCTs.

CONCLUSION

The starting point of understanding RCT failure is typically an assumption that failure is likely to mean that the underlying causal theory is faulty, or the implementation is at fault. Stame's (2010) work adds potential for programme and method-related faults. These explanations focus on the programme design and delivery, as well as its evaluation. However, taking a broader perspective on the context within which programmes and RCTs take place allows us to see that there is much more in play that needs to be taken into account in our explanations of RCT failure (or success). The political and funding backdrop, in particular, are often obscure in evaluations of RCTs; yet, as can be seen in the case of the Integrating English study, these can play a crucial role in helping determine—even at the outset—the likelihood of success or failure.

In the case of Integrating English, the model presented in this paper has demonstrated how a range of issues from the funder's remit, through the evaluation design and programme operation, to the wider political context all contributed to failure. Yet as indicated in the previous section, the outcome led to the demise of the programme, at least as far as the EEF is concerned. The array of issues identified worked together in a systematic way to create—with hindsight—a system context that made success extremely unlikely. This is likely to have wider applicability for other RCT studies, especially for cross-curricular programmes. We might characterise this not as a failure of the programme or the evaluation, but rather as a failure of the system to create the conditions for potential success.

The purpose of this analysis is not to play down or explain away RCT failure; rather, it is to help researchers—and funders, commissioners, practitioners and policymakers—to understand this failure better, and therefore make better decisions. It should be noted, in fact, that these elements are of use in better understanding the results of RCTs whether the programme indicates impact or not.

The analysis presented in this paper helps make plain that the veneer of objectivity surrounding RCTs is illusory. While it is possible to demonstrate, logically, the robustness of statistical relationships between measures indicating an effect on a population, how, where and what is measured are choices dependent on theory, prior evidence, how data is defined and gathered, some speculation and hypothesising and a subjective standpoint situated within a socio-historical moment. That is, RCTs, like other methods, are subjective in their focus, their means of implementation and the measures used (Gillborn et al., 2018). For instance, the high-stakes SATs tests in the United Kingdom, used as an impact measure in many EEF trials, are not an objective measure of success in primary education; the SATs for English promote a view of 'standard English' which works to enforce 'prescriptive linguistic ideologies' (Cushing, 2021, p. 599). It is only with this understanding—that the commission, design, operation and interpretation of RCTs are all matters of choice—that we can fully appreciate the value of the evidence that an RCT can provide. Reflecting on our methods can help us realise that the way we construct knowledge through RCTs is itself a choice of our discipline (Burnett & Coldwell, 2021; Law & Ruppert, 2013), which enacts the current values of policymakers, funders and educators (Wacquant, 2022). Fully appreciating the choices that produce RCT failure and success will help us better interpret the evidence they produce.

Further, this analysis suggests that policymakers and funders should think again about the response to RCT failure (and indeed success). RCTs do not fail simply due to problems with underlying programme theory or its implementation; the failure, as we note, can reflect much broader issues that funders need to consider. This paper provides a starting

point, in the context-informed explanatory framework (Table 2), to help identify these issues. Naturally, whilst some of the issues identified here could be rectified, many of them are well beyond the scope of the funder, evaluator or developer to change. In this sense, the case study described here not only contributes to our understanding of RCTs and their success or failure, but also adds further evidence to the body of research demonstrating the consequences of the narrow and test-based accountability and regulatory systems (Acosta et al., 2020; Au, 2007; Brill et al., 2018; Mausethagen, 2013) in place in education in England and across the world.

ACKNOWLEDGEMENTS

The authors are most grateful to their colleagues at the Sheffield Institute of Education who were engaged on the evaluation of Integrating English, to Professor Emerita Bronwen Maxwell and Professor Mark Boylan for timely and valuable observations, conversations and contributions, and to two anonymous reviewers for very helpful comments that we believe have produced a stronger piece of writing. Any errors or omissions remain the responsibility of the authors.

FUNDING INFORMATION

This study was funded by the Education Endowment Foundation, the Bell Foundation and Unbound Philanthropy.

CONFLICT OF INTEREST STATEMENT

There are no conflicts of interest to declare. For transparency, the authors declare their part in the independent evaluation team for the RCT and accompanying IPE.

DATA AVAILABILITY STATEMENT

No new data was generated for this publication.

ETHICS STATEMENT

The study design for the research data drawn upon in this paper was independently reviewed and approved by the Sheffield Hallam University Ethics Committee in 2016 (Ref. AM/RKT/297-CUL; see www.shu.ac.uk/research/excellence/ethics-and-integrity), based on data previously published (https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/integrating-english) under International Standard Randomised Controlled Trial No. ISRCTN44415239.

ORCID

Mike Coldwell https://orcid.org/0000-0002-0829-5523
Nick Moore https://orcid.org/0000-0002-7385-3077

REFERENCES

Acosta, S., Garza, T., Hsu, H. Y., Goodson, P., Padrón, Y., Goltz, H. H., & Johnston, A. (2020). The accountability culture: A systematic review of high-stakes testing and English learners in the United States during No Child Left Behind. *Educational Psychology Review*, *32*(2), 327–352. https://doi.org/10.1007/s10648-019-09511-2

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*(5), 258–267. https://doi.org/10.3102/0013189X07306523

Boylan, M., Coldwell, M., Maxwell, B., & Jordan, J. (2018). Rethinking models of professional learning as tools: A conceptual analysis to inform research and practice. *Professional Development in Education*, *44*(1), 120–139. https://doi.org/10.1080/19415257.2017.1306789

Breuer, E., Lee, L., De Silva, M., & Lund, C. (2015). Using theory of change to design and evaluate public health interventions: A systematic review. *Implementation Science*, *11*(1), 63. https://doi.org/10.1186/s13012-016-0422-6

Brill, F., Grayson, H., Kuhn, L., & O'Donnell, S. (2018). *What impact does accountability have on curriculum, standards and engagement in education? A literature review*. National Foundation for Educational Research.

Burnett, C., & Coldwell, M. (2021). Randomised controlled trials and the interventionisation of education. *Oxford Review of Education*, *47*(4), 423–438. https://doi.org/10.1080/03054985.2020.1856060

Coldwell, M. (2019). Reconsidering context: Six underlying features of context to improve learning from evaluation. *Evaluation*, *25*(1), 99–117. https://doi.org/10.1177/1356389018803234

Coldwell, M., & Maxwell, B. (2018). Using evidence-informed logic models to bridge methods in educational evaluation. *The Review of Education*, *6*(3), 267–300. https://doi.org/10.1002/rev3.3151

Connell, J., & Kubisch, A. (1998). Applying a theory of change approach to the evaluation of comprehensive community initiatives: Progress, prospects and problems. In K. Fulbright-Anderson, A. Kubisch, & J. Connell (Eds.), *New approaches to evaluating community initiatives, Vol. 2: Theory, measurement and analysis*. Aspen Institute.

Culliney, M., Moore, N., Coldwell, M., & Demack, S. (2019). *Integrating English – evaluation report*. Education Endowment Foundation.

Cushing, I. (2021). Grammar tests, de facto policy and pedagogical coercion in England's primary schools. *Language Policy*, *20*, 599–622. https://doi.org/10.1007/s10993-020-09571-z

Custance, B., Dare, B., & Polias, J. (2012). *Teaching ESL students in mainstream classrooms: Language in learning across the curriculum*. Lexis Education.

De Costa, P., & Qin, K. (2015). English language education in the United States: Past, present and future issues. In L.T. Wong & A. Dubey-Jhaveri (Eds.), *English language education in a global world* (pp. 229–238). Nova Science.

Eccles, M., & Mittman, B. (2006). Editorial: Welcome to *Implementation Science*. *Implementation Science*, *1*(1), 1–3.

EEF. (2019). *Annual report 2019*. EEF https://d2tic4wvo1iusb.cloudfront.net/production/documents/annual-reports/EEF_2019_Annual_Report_-_printable.pdf?v=1688735468

EEF. (2022a). *Education Endowment Foundation projects*. EEF https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects

EEF. (2022b). *Education Endowment Foundation: Who we are*. EEF https://educationendowmentfoundation.org.uk/about-us/who-are-we

EEF. (2023a). *Pipeline of EEF trials*. EEF https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/process-and-people/pipeline-of-eef-trials

EEF. (2023b). *Integrating English EEF*. EEF https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/integrating-english

Fang, Z., & Schleppegrell, M. (2010). Disciplinary literacies across content areas: Supporting secondary reading through functional language analysis. *Journal of Adolescent and Adult Literacy*, *53*(7), 587–597. https://doi.org/10.1598/JAAL.53.7.6

Gillborn, D., Warmington, P., & Demack, S. (2018). QuantCrit: Education, policy, 'big data' and principles for a critical race theory of statistics. *Race Ethnicity and Education*, *21*(2), 158–179. https://doi.org/10.1080/13613324.2017.1377417

Greany, T., & Higham, R. (2018). *Hierarchy, markets and networks: Analysing the 'self-improving school-led system' agenda in England and the implications for schools*. UCL IoE Press https://discovery.ucl.ac.uk/id/eprint/10053501/1/Hierarchy%20Markets%20and%20Networks%20FINAL.pdf

Greenhalgh, J., & Manzano, A. (2021). Understanding 'context' in realist evaluation and synthesis. *International Journal of Social Research Methodology*, *25*, 583–595. https://doi.org/10.1080/13645579.2021.1918484

Halliday, M. A. K. (1993). Towards a language-based theory of learning. *Linguistics and Education*, *5*, 93–116. https://doi.org/10.1016/0898-5898(93)90026-7

Haynes, L., Service, O., Goldacre, B., & Torgerson, D. (2012). *Test, learn, adapt: Developing public policy with randomised control trials*. Cabinet Office Behavioural Insights Team https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/62529/TLA-1906126.pdf

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

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2019). *Implementation and process evaluation (IPE) for interventions in education settings: An introductory handbook*. EEF https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/IPE_Handbook.pdf

Hutchinson, J. (2018). *Educational outcomes of children with English as an additional language*. The Bell Foundation www.bell-foundation.org.uk/app/uploads/2018/02/Educational-Outcomes-of-Children-with-EAL.pdf

Jacob, R. T., Doolittle, F., Kemple, J., & Somers, M.-A. (2019). A framework for learning from null results. *Educational Researcher*, *48*(9), 580–589. https://doi.org/10.3102/0013189X19891955

Jacobson, M. J., Levin, J. A., & Kapur, M. (2019). Education as a complex system: Conceptual and methodological implications. *Educational Researcher*, *48*(2), 112–119. https://doi.org/10.3102/0013189X19826958

Law, J., & Ruppert, E. (2013). The social life of methods: Devices. *Journal of Cultural Economy*, *6*(3), 229–240. https://doi.org/10.1080/17530350.2013.812042
Lendrum, A., & Humphrey, N. (2012). The importance of studying the implementation of interventions in school settings. *Oxford Review of Education*, *38*(5), 635–652. https://doi.org/10.1080/03054985.2012.734800
Long, R., & Danechi, S. (2022). *Teacher recruitment and retention in England*. House of Commons Library https://researchbriefings.files.parliament.uk/documents/CBP-7222/CBP-7222.pdf
Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, *48*(3), 158–166. https://doi.org/10.3102/0013189X19832850
Mausethagen, S. (2013). A research review of the impact of accountability policies on teachers' workplace relations. *Educational Research Review*, *9*, 16–33. https://doi.org/10.1016/j.edurev.2012.12.001
Maxwell, B., Sharples, J., & Coldwell, M. (2022). Developing a systems-based approach to research use in education. *The Review of Education*, *10*(3), e3368. https://doi.org/10.1002/rev3.3368
Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. https://doi.org/10.3102/0013189X018002005
Murphy, V. A., & Unthiah, A. (2015). *A systematic review of intervention research examining English language and literacy development in children with English as an additional language (EAL)*. EEF https://d2tic4wvo1iusb.cloudfront.net/documents/guidance/EAL_Systematic_review.pdf?v=1629122270
Nevill, C. (2019). *EEF blog: Randomised control trials – 3 good things, 3 bad things, and 5 top tips*. EEF https://educationendowmentfoundation.org.uk/news/eef-blog-randomised-controlled-trials-or-how-to-train-your-dragon
Nielsen, K., Fredslund, H., Christensen, K. B., & Albertsen, K. (2006). Success or failure? Interpreting and understanding the impact of interventions in four similar worksites. *Work and Stress*, *20*(3), 272–287. https://doi.org/10.1080/02678370601022688
Nielsen, S. B., Lemire, S., & Tangsig, S. (2021). Unpacking context in realist evaluations: Findings from a comprehensive review. *Evaluation*, *28*(1), 91–112. https://doi.org/10.1177/13563890211053032
Opfer, V. D., & Pedder, D. (2011). Conceptualizing teacher professional learning. *Review of Educational Research*, *81*(3), 376–407. https://doi.org/10.3102/0034654311413609
Pawson, R. (2013). *The science of evaluation: A realist manifesto*. SAGE.
Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. SAGE.
Rog, D. J. (2012). When background becomes foreground: Toward context-sensitive evaluation practice. In D. J. Rog, J. L. Fitzpatrick, & R. F. Conner (Eds.), *Context: A framework for its influence on evaluation practice* (Vol. *135*, pp. 25–40). New Directions for Evaluation.
Rogers, P. (2014). *Theory of change, methodological briefs: Impact evaluation 2*. UNICEF Office of Research.
Rogers, P. J. (2011). Implications of complicated and complex characteristics for key tasks in evaluation. In K. Forss, M. Marra, & R. Schwartz (Eds.), *Evaluating the complex – attribution, contribution and beyond* (pp. 33–52). Routledge.
Rogers, P. J., & Weiss, C. H. (2007). Theory-based evaluation: Reflections ten years on. Theory-based evaluation: Past, present, and future. *New Directions for Evaluation*, *114*, 63–81. https://doi.org/10.1002/ev.225
Rose, D., & Martin, J. R. (2012). *Learning to write, reading to learn: Genre, knowledge and pedagogy of the Sydney school*. Equinox.
Smith, B., Morris, S., & Armitage, H. (2021). *The effects of using examination grade as a primary outcome in education trials to evaluate school-based interventions*. EEF https://educationendowmentfoundation.org.uk/public/files/Publications/The_effects_of_using_examination_grade_as_a_primary_outcome_in_education_trials.pdf
Stame, N. (2010). What doesn't work? Three failures, many answers. *Evaluation*, *16*(4), 371–387. https://doi.org/10.1177/1356389010381914
Strand, S., Malmberg, L., & Hall, J. (2015). *English as an additional language (EAL) and educational achievement in England: An analysis of the National Pupil Database*. EEF https://d2tic4wvo1iusb.cloudfront.net/documents/guidance/EAL_and_educational_achievement__Prof_S_Strand.pdf?v=1629122217
Strom, K. J., Mills, T., & Abrams, L. (Eds.). (2023). *Non-linear perspectives on teacher development – complexity in professional learning and practice*. Routledge.
Supovitz, J. (2009). Can high stakes testing leverage educational improvement? Prospects from the last decade of testing and accountability reform. *Journal of Educational Change*, *10*, 211–227. https://doi.org/10.1007/s10833-009-9105-2
Thompson, G., Lingard, B., & Ball, S. J. (2021). 'Indentured autonomy': Headteachers and academisation policy in Northern England. *Journal of Educational Administration and History*, *53*(3–4), 215–232. https://doi.org/10.1080/00220620.2020.1850433
Unsworth, L., Tytler, R., Fenwick, L., Humphrey, S., Chandler, P., Herrington, M., & Pham, L. (2022). *Multimodal literacy in school science – transdisciplinary perspectives on theory, research and pedagogy*. Routledge.
Veel, R. (1997). Learning how to mean – scientifically speaking: Apprenticeship into scientific discourse in the secondary school. In F. Christie & J. R. Martin (Eds.), *Genre and institutions: Social processes in the workplace and school* (pp. 161–195). Pinter.

Wacquant, L. (2022). Epistemic bandwagons, speculation, and turnkeys: Some lessons from the tale of the urban 'underclass'. *Thesis Eleven*, *173*(1), 82–92. https://doi.org/10.1177/07255136221121705

What Works Network. (2019). *Guidance: What Works Network*. www.gov.uk/guidance/what-works-network

Wilkins, A. (2017). Rescaling the local: Multi-academy trusts, private monopoly and statecraft in England. *Journal of Educational Administration and History*, *49*(2), 171–185. https://doi.org/10.1080/00220620.2017.1284769

Wrigley, T. (2018). The power of 'evidence': Reliable science or a set of blunt tools? *British Educational Research Journal*, *44*(3), 359–376. https://doi.org/10.1002/berj.3338

How to cite this article: Coldwell, M. & Moore, N. (2024). Learning from failure: A context-informed perspective on RCTs. *British Educational Research Journal*, *00*, 1–21. https://doi.org/10.1002/berj.3960

APPENDIX A

THE CONTEXT-INFORMED EXPLANATORY FRAMEWORK APPLIED TO THE INTEGRATING ENGLISH RCT

Context category/ phase	Contexts (developed from Rog, 2012)	Forms of failure (Stame, 2010)			
		Theory	Programme	Method	Implementation
Backdrop	Problem			*Nature and composition of EAL classes in sample*	*Incentivisation of participation in place of viable alternatives for schools*
	Decision-making			*EEF and What Works Centres favour RCTs*	
Design	Causal theory	**Limited evidence of complex change process**			
	Intervention		**Choice to intervene in primary rather than secondary schools**		
	Evaluation			**Use of possibly inappropriate measures of change.** *Little chance of seeing change in high-stakes tests like SATS*	
Operation	Teaching				**Low return rate of some measures of fidelity. Little evidence of training producing change in planning or in classrooms**
	Broader environment			*Spread over two school years and large variation in subjects*	
Interpretation	Consequential	*Unclear what failed—Integrating English trial, EAL teaching, SFL—as no further funding available*		*Influence of RCTs as evidence for What Works Centres*	

Contributory factors in *italics*; critical factors in **bold**.