Research Paper

ECNU Review of Education
2021, Vol. 4(1) 46–64

DOI: 10.1177/2096531120913039
journals.sagepub.com/home/roe

Working Out What Works: The Case of the Education Endowment Foundation in England

Triin Edovald and Camilla Nevill
The Education Endowment Foundation

Abstract

Purpose: This article gives an overview of the successes and lessons learned to date of the Education Endowment Foundation (EEF), one of the leading organizations of the What Works movement.

Design/Approach/Methods: Starting with its history, this article covers salient components of the EEF's unique journey including lessons learned and challenges in evidence generation.

Findings: The EEF has demonstrated that it is feasible to rapidly expand the use of school-based randomized controlled trials (RCTs) in a country context, set high standards for research independence, transparency, and design, and generate new evidence on what works. Challenges include the need to consider alternative designs to RCTs to answer a range of practice-relevant questions, how to best test interventions at scale, and how study findings are reported and interpreted.

Originality/Value: This article addresses some of the key components required for the success of What Works organizations globally.

Keywords

Education, evaluation, evidence, impact, randomized controlled trials, teaching

Date received: 22 October 2019; accepted: 18 January 2020

Corresponding author:
Triin Edovald, The Education Endowment Foundation, 5th Floor, Millbank Tower, 21–24 Millbank, London SW1P 4QP, UK.
Email: triin.edovald@eefoundation.org.uk

Introduction

The articles in this issue focus on a general introduction to evidence-based educational reform in different countries and the effectiveness of evidence-based education reforms. There have been various identifiable waves in the process of evidence revolution in education with the 2010s characterized by attempts to institutionalize the use of evidence through the emergence of knowledge brokering agencies, most notably the What Works movement in the U.S. and the UK (White, 2019). This article focuses on the case of the Education Endowment Foundation (EEF) in England and presents an internal perspective on the work of the EEF in generating, documenting, and promoting the use of high-quality evidence and evaluation to inform teaching and other school practices, for transfer and possible adaptation in other contexts. The article gives a brief overview of the EEF's history, discusses the successes and lessons learned so far, followed by a discussion on the challenges faced primarily from the perspective of evidence generation.

A brief history of the EEF

Inspired by the Obama administration's Race to the Top initiative in the U.S. (U.S. Department of Education, 2009), the UK Secretary of State for Education, Michael Gove announced in late 2010 plans to establish an EEF to help raise standards in challenging schools in England (Department for Education and The RT Hon Michael Gove MP, 2010). The EEF was founded in 2011 by a lead charity, The Sutton Trust, in partnership with Impetus, with a £125 million founding grant from the Department for Education. They were selected following an open competition which attracted interest from 14 organizations. The EEF was envisaged to have a 15-year life span. In addition to receiving the founding grant, the EEF has set itself the goal of securing additional investment to enable it to award over £200 million in supporting the development, delivery, and rigorous evaluation of programs over its life span (The Education Endowment Foundation [EEF], 2012a).

The EEF is governed by an independent Board of Trustees, nominated by the founding partners and Chaired by Sir Peter Lampl. The Board and the executive team are guided by two key advisory bodies: an Advisory Board of experts from education, public policy, and business; and an Evaluation Advisory Group (EAG). The EAG provides critical guidance on evaluation methodologies and best practice in evidence generation. The charity is also supported by a number of legal and professional services firms, offering pro bono advice. Importantly, the EEF is independent of government, but maintains strong and collaborative working relationships with a number of Ministries, principally the Department for Education.

In March 2013, the EEF and the Sutton Trust were jointly designated by the Government as the What Works Clearinghouse (WWC) for Education. The WWC network is made up of nine independent WWCs, three affiliate members, and two associate members (Cabinet Office, 2019). Together these centers cover policy areas which account for more than £250 billion of

public spending. What sets WWCs apart from standard research institutions is that the centers are committed to increasing both the supply of and demand for evidence in their policy area, and their output is tailored to the needs of their primary audiences (Cabinet Office, 2013; Cabinet Office and HM Treasury, 2018).

The Government's WWC network represents the political hegemony of the What Works movement, which is largely built on the rise of impact evaluations (particularly randomized controlled trials [RCTs]) since the early 2000s and the increased production of systematic reviews over the last 10 years (White, 2019). These developments have been paralleled in the UK with the emphasis placed on evidence-based policy and practice by the incoming New Labour Government in 1997 and taken forward through successive Labour administrations, the following Conservative/Liberal Democrat Coalition Government 2011–2015 and the Conservative Government thereafter (see also Connolly et al., 2017).

Mission, aims, and key strands of work of the EEF

The EEF is an independent charity dedicated to breaking the link between family income and educational achievement in England. It aims to raise the attainment of 3 to 18-year-olds, particularly those facing disadvantage, develop their essential life skills, and prepare young people for the world of work and further study.

The EEF is dedicated to achieving its aims by

1. synthesizing the best available evidence in user-friendly language for senior leaders and teachers in schools and translating this into resources that include summaries and practical tools and that are designed to improve practice and boost learning (e.g., Quigley & Coleman, 2019; van Poortvliet et al., 2018; The Teaching and Learning Toolkit available at https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit);
2. generating evidence of what works to improve teaching and learning, by funding high-quality, independent evaluations of promising programs and interventions using predominantly RCT designs; and
3. supporting educational institutions ranging from early years' settings to post-16 settings across England by promoting the use of evidence to inform practice to maximize the benefits for children and young people. To do so the EEF works in partnership with a network of 32 Research Schools and 8 Associate Research Schools across the country to support the use of evidence to improve teaching practice (see www.researchschool.org.uk/ for further info).

Currently, there are 20,217 state-funded schools, attended by just over 8 million pupils, in England (Department for Education, 2019a) with 453,400 full-time equivalent teachers working in state schools (Department for Education, 2019b). As of March 2020, just over 14,000 schools

and over 1.58 million children and young people were involved in EEF-funded studies (of which over 150 are RCTs). A recent systematic review found that there have been 1,017 unique RCTs in education since 1980 and of these, 799 have been produced in the last 10 years (Connolly et al., 2018). Only 25% of education trials identified in this review included more than 1,000 participants (Connolly et al., 2018), whereas over 70% of EEF RCTs include more than 1,000 participants, with the average size being over 8,000 participants. Therefore, the EEF is one of the leading funders of RCTs in education globally and has commissioned approximately 19% of all known trials in the last 10 years, and some of the largest.

While the EEF's work has a predominantly domestic focus, its approach to generating and using evidence to improve teaching and learning is internationally relevant. Furthermore, more countries becoming involved in this endeavor will support the EEF's core mission to boost attainment for disadvantaged children and young people in England. The EEF has quickly established itself as a world-leading organization in evidence generation and supporting teachers to put research to good use, which has triggered the development of a number of global partnerships since 2014 spanning Australasia to Latin America. In 2018, the EEF launched a 5-year project "Building a Global Evidence Ecosystem for Teaching" in partnership with the BHP Foundation as part of its Global Education Equity Program. This project will enable the EEF to take its work to scale, supporting more partners in more countries to generate evidence that senior leaders and teachers could use to make evidence-informed decisions to support school improvement.

Evidence generation: Key successes and lessons learned

Feasibility of large-scale commissioning of sizable school-based education evaluations in a single country context

The EEF has demonstrated that it is possible to commission many, relatively large, education evaluations, mostly RCTs, over a short period of time (8 years). Prior to the EEF being set up in 2011, few large-scale pragmatic RCTs had been conducted in English schools (Styles & Torgerson, 2019). There was some resistance to the use of RCTs in the UK among parts of the education research community (Oakley, 2006) and common objections included the argument that randomization itself is unethical and the perception that participants will think the randomization is unethical and refuse to participate (Hutchinson & Styles, 2010). Indeed, it was believed that one of the main challenges the EEF would face would be persuading schools to take part, yet by 2019, it had successfully recruited more than half the schools in England to its evaluations (Nevill, 2019a).

Early EEF trials suffered from relatively high attrition, with an average of 24% of pupils dropping out between 2011 and 2012 (Dawson et al., 2017). Since then, the EEF has introduced a number of strategies to recruit and retain schools. For example, the EEF has learned the value of

communicating the benefits of RCTs to schools through recruitment events and documents clearly explaining the evaluation design and the schools' responsibilities (Dawson et al., 2017), strategies which are supported by the literature on intrinsic motivation (Tirole & Benabou, 2006). Extrinsic rewards can also have a role in increasing intrinsic motivation (Muralidharan & Sundararaman, 2011). The EEF offers financial incentives, particularly to control schools, and where the data collection burden is onerous. The EEF has also recognized schools' contribution through letters of thanks and certificates designating them "EEF research partner schools."

These efforts have culminated in the EEF developing and publishing guidance on recruitment and retention for delivery partners (The EEF, 2019a). Recruitment has become easier over time, as schools have become familiar with EEF's work, and the education system has recognized the value of high-quality research (Cullinane, 2018).

The EEF has set high standards for research independence, transparency, and design

With £125 million available at the outset, the EEF was able to set the agenda regarding the type of research and evaluation it would fund. It began with the intention of only commissioning RCTs as "gold standard" evaluation designs with respect to minimizing selection bias. For example, the U.S. Government's What Works Clearinghouse (WWC, 2017) gives its highest rating only to well-implemented RCTs. Yet, not all RCT designs provide useful evidence for schools and there are many aspects of both design and delivery that are debated (Ginsburg & Smith, 2016).

This section discusses some salient aspects of EEF's journey in the development of its quality standards and expectations regarding gold standard evidence generation. The EEF has taken a collaborative approach to develop these, via consultation with the UK and international research organizations, relevant experts, and its Evaluation Advisory Group (see above on governance structure). For example, the EEF hosts an annual conference and analysis workshop for the members of its Panel of Evaluators to discuss challenges and solutions in education evaluation.

Independent evaluation. Conflict of interest in summative evaluation is a recognized problem across many fields. In medicine, drug trials are often funded by drug companies, leading to accusations of reporting bias and negative findings being withheld (Goldacre & Heneghan, 2014). Similarly, in education, there is a risk of bias from developers who exert a strong influence over the study and have a personal or financial interest in a positive result. In particular, "as one reaches the stage of summative evaluation, there are clear concerns about bias when an evaluator is too closely affiliated with the design team" (National Research Council, 2004, p. 61). Yet historically such developer-led evaluations have been common (Ginsburg & Smith, 2016).

The EEF recognized the benefits of setting a precedent for commissioning "independent" evaluations from the outset and needed to quickly clarify what it meant by that. The approach the

EEF took was to appoint, through a competitive tendering process, a Panel of Evaluators, namely research organizations with expertise in education evaluation, who would compete to evaluate programs, and subsequently be partnered with developers. The EEF has separate grant agreements with the evaluator and developer and acts as a mediator in discussions about evaluation design, implementation, and analysis. The EEF's standards for independent evaluation clarify that it expects design decisions to be made collaboratively, but that randomization, primary outcome data collection, analysis, and reporting should always be conducted by the evaluator (The EEF, 2017a).

The approach chosen by the EEF is unusual and relatively extreme in the way that it separates evaluators' and developers' financial and personal interests. For example, a comparable organization to the EEF is the U.S. Department of Education's Institute for Education Sciences (IES) and its evaluation arm; the National Centre for Education Evaluation and Regional Assistance (NCEE), which conducts and supports large-scale evaluations of education programs. The IES was established by the Education Sciences Reform Act (ESRA) of 2002.

ESRA requires that the NCEE conduct independent evaluations by "awarding evaluation contracts competitively to experts external to the Department who are free from conflicts of interest" (IES, 2017). Yet, to meet this requirement, grantees are expected to select the evaluator themselves, name them in their grant application, and are responsible for administering the grant funds. In a report of 65 Investing in Innovation (i3) education evaluations, supported by the IES NCEE,[1] it was found that 97% were independent, as defined by having at least one outcome that was collected, analyzed, and reported by the independent evaluator (Boulay et al., 2018). Independent evaluators are usually from a different organization to the grantee or developer, but there appear to be no checks regarding evaluator's conflict of interest, unlike in the EEF commissioning model.

One consequence of this approach may be that there are fewer RCTs and more quasi-experimental designs (QEDs) commissioned. A QED approach is often preferred by developers for practical reasons, and evaluators may feel compelled to agree if they are paid by the developers. Of the 19 i3 impact evaluations supported by NCEE,[2] 13 (just over two thirds) were reported to meet the WWC standards without reservations (Boulay et al., 2018), meaning a randomized design without high attrition, defined as less than 55% overall (WWC, 2017). It is not possible to directly compare the WWC standards with the EEF's padlock rating that is used to assess the security of the primary impact result (see the section on challenges below). However, of the 95 published impact evaluation reports to date, 89% were based on an RCT design and 85% achieved three or more padlocks, meaning that they had an overall attrition rate of less than 30%.

The approach taken by the EEF was controversial, there has been resistance in some parts of the academic community, and several projects have fallen through because the developers would not

agree to the design proposed by the independent evaluator. But it has been successful in achieving high-quality results and minimizing bias.

Reporting transparency. Publication bias is a recognized problem whereby authors and publishers favor positive findings. An analysis of social science research found that those with strong results are 40 percentage points more likely to be published than null results (Franco et al., 2014). For this reason and to minimize selective reporting (i.e., the bias that derives from the exclusion of negative or undesirable results), the EEF requires a prespecified protocol and statistical analysis plan for every trial to be published on its website and the trial registered on ISRCTN registry,[3] a primary clinical trial registry. The first EEF protocol and reporting templates were published in 2013, based on CONSORT standards (Shulz et al., 2010), and have been updated since to reflect changing standards (Montgomery et al., 2018). All EEF's findings are published, whatever the result (Nevill, 2016).

Despite these measures being taken by the EEF, the risk of publication bias still exists. There have been several examples where the developer has chosen to separately publish journal articles that present a more positive picture, in response to EEF's report of a null result (e.g., Burgess et al., 2019). The EEF attempts to minimize this risk by encouraging evaluators and developers to enter into a publication agreement and requesting to see publications before they are submitted. But in practice, once the evaluation is complete, it has little influence.

Data archiving and reproducibility. There have been concerns over the last decade regarding the replicability of scientific findings (Open Science Collaborative, 2015). The EEF knew it would be generating large amounts of powerful RCT data, so with urgency in 2012, it set up a data archive to enable it to check the reproducibility of evaluator estimates, track long-term outcomes, and support secondary research and reanalysis across trials. Education researchers in the UK benefit from having access to high-quality census data on pupils and schools, including outcomes at the end of primary and secondary school, compiled by the UK Department for Education (FFT Education Datalab, 2018). This is called the National Pupil Database (NPD). The EEF data archive is the first of its kind to collect pupil-level data from many large-scale education RCTs in one place, link this data set to longitudinal outcomes (in the NPD), and make it accessible for further research.

It has led to important methodological innovations. For example, the reanalysis of 17 early EEF trials using four analytical models enabled EEF research partners at Durham University to reveal the extent of variation in effect size estimates that occurs as a result of analysis choice (Xiao et al., 2016) and led to the EEF publishing the first version of its statistical analysis guidance to increase the comparability of trial results, which has been updated three times since (The EEF, 2018). It has also been used to examine the theoretical and empirical implications of accounting for clustering at

the class level (Demack, 2019) and to explore using standard deviation (SD) as an outcome of an intervention (Tymms & Kasim, 2018).

Currently, the archive holds data for 105 completed studies that the EEF has commissioned. As the number of trials hosted within the EEF archive grows, this powerful data set provides increasing potential for understanding what works for different types of schools and pupils.

Evaluation design. There is much that could be written about EEF's journey with respect to the designs that it commissions, but this article will focus on two aspects: implementation and process evaluation (IPE) and measurement of outcomes.

IPE. Some early EEF trials suffered because they lacked high-quality IPE which meant that they were unable to explain the causal processes underlying the results or describe implementation (Morris et al., 2016). This is not unusual in education, with only 38% of 1,017 education trials identified between 1980 and 2016, including a process evaluation component (Connolly et al., 2018). But EEF's early approach to design was limiting because education programs are complex and there is much to be learned regarding implementation and causal mechanisms. The British Medical Research Council's evaluation recommendations for complex interventions specify that "a good theoretical understanding is needed of how the intervention causes change" and "lack of effect may reflect implementation failure (or teething problems) rather than genuine ineffectiveness" (Craig et al., 2008, p. 980). For this reason, in 2014, the EEF commissioned a literature review by Manchester University of IPE for education interventions (Humphrey et al., 2016) which informed guidance highlighting the importance of a detailed intervention description (Hoffman et al., 2014) and high-quality data on implementation (Durlak & Dupre, 2008), compliance, control group activity, causal mechanisms, and cost (Dawson et al., 2017). Since then the EEF has commissioned a review of methods for evaluating complex education programs (Anders et al., 2017) and published revised guidance with the aim of improving theory-testing, integration of impact and IPE, prespecification, and the measurement of compliance (The EEF, 2019b).

For commissioning bodies like the EEF, there is a difficult balance to strike between the desire to evidence many hypothesized elements of the intervention's theory of change and the practical risk of overburdening participants and cost considerations. There are several examples of the EEF commissioning multiarmed trials (e.g., Lord et al., 2017; McNally et al., 2018), and 90% of EEF trials include the measurement of at least one secondary outcome or mechanism of change (Nevill, 2019a). However, there is still some distance to travel before the EEF is regularly commissioning all aspects of "realist" trials, for example, that

> examine the effects of intervention components separately and in combination; using multi-arm studies and factorial trials; explore mechanisms of change, for example analysing how pathway variables

> mediate intervention effects; use multiple trials across contexts to test how intervention effects vary with contexts. (Bonell et al., 2012, p. 2299)

Measurement of outcomes. A study may have the most precise analysis but without psychometrically reliable and valid measurement instruments, RCT results have little meaning. At the outset, the EEF recognized the risk associated with using outcomes too closely aligned to the treatment (Ginsburg & Smith, 2016; WWC, 2017) resulting in greatly inflated effect sizes (Cheung & Slavin, 2016). Early in 2012, the EEF published strict guidance on test selection for the primary outcome in its evaluations, the main criteria being that it must have broad external validity and be highly correlated with performance in national high stake assessments (The EEF, 2012b). The EEF also uses outcomes from the NPD where possible. Further discussion on the approach the EEF has taken to address measurement, attrition, and timing can be found in Dawson et al. (2017).

There are many available commercial standardized tests in the UK, with some being widely used by schools and researchers. Yet, the psychometric properties of these tests are not well reported. Some companies do report measures such as internal test-item consistency (e.g., Cronbach's α), but studies of validity and reliability are less frequent. Recent analysis of archived data has shown that many of these assessments offer only moderate predictive validity (Allen et al., 2018) and some evaluations have suffered from floor and ceiling effects (e.g., Hodgen et al., 2019). In 2014, the EEF expanded its remit to include the early years and non-attainment outcomes such as self-control and resilience and commissioned literature reviews to inform databases of available measures in these areas (e.g., Wigelsworth, 2017). The EEF is now commissioning a systematic review of available attainment measures to fill this gap but would have benefited from doing so earlier. This is a useful lesson for similar organizations wanting to fund large numbers of evaluations with similar outcomes.

The EEF has generated evidence about what does and does not work

As a result of the efforts described above, the EEF has generated a large body of evidence that helps to identify what does and does not work in English schools. For example, EEF research has clarified the need to better deploy teaching assistants and the importance of early years' approaches (The EEF, 2017b). It has also generated important knowledge about what does not work and that schools should be wary about expecting large returns from popular approaches such as "lesson study" and "growth mindset" (Foliano et al., 2019; Murphy et al., 2017). Much of what has been learned has not been from the headline finding and schools find information on implementation useful (Quigley, 2019).

Synthesis is essential in order to establish the external validity of findings (Shadish et al., 2002). For this reason, EEF trials are part of a rich tapestry of evidence generated by the EEF including the

Teaching and Learning Toolkit (Higgins et al., 2015) and Early Years Toolkit, which is currently based on many meta-analyses (Higgins et al., 2013), and EEF's guidance reports which provide practical, evidence-based guidance for teachers on a range of high-priority issues, based on the best available evidence. The EEF Toolkits present over 30 approaches to improving teaching and learning, each summarized in terms of its average impact on attainment, its cost, and the strength of the evidence supporting it (Higgins et al., 2015). Recent guidance reports include recommendations on improving social and emotional learning in primary schools (van Poortvliet et al., 2019) and improving literacy in secondary schools.

The meta-analyses in the Teaching the Learning Toolkit often combine studies of varying quality, on different ages, subjects, and even countries. This is why the EEF has commissioned the EEF Education Database, a major initiative involving scores of coders coding the estimated 10,000 individual studies within the Toolkit. The ultimate aspiration is to create a "live" database where all education studies carried out globally can be included as their results become available.

Key challenges faced and opportunities for future

While the EEF has had the opportunity to celebrate many milestones on its evidence synthesis, generation, and mobilization journey, it has also faced several challenges, which have pushed it to adapt and innovate.

Common RCT designs are not always suited to answering some kinds of questions of importance to schools and teachers

The EEF typically commissions relatively large-scale trials using school-level randomization. However, the EEF has learned that it can sometimes be hard to determine in advance whether an RCT is the most feasible way to evaluate an intervention. Uncertainties can emerge in the design and implementation stages regarding various trial aspects such as a number of participants, likely intervention effects, and costs (Edovald & Firpo, 2016). Furthermore, sometimes an RCT design is not acceptable to participants (e.g., Sutherland et al., 2017). Also, some interventions or questions lend themselves more readily to RCTs than others. This section describes two new strands of work that the EEF has recently introduced in response to the challenges it has faced in evaluating certain relevant questions in school settings.

School choices. Successful RCT designs rely on participants that are willing to be randomized. The EEF has managed to recruit more than half the schools in England to participate in its RCTs, but there are some choices that schools are not willing to be randomized to. For example, an EEF-funded RCT of mixed ability grouping versus setting and streaming based on ability failed to recruit schools (Roy et al., 2014), as did a trial that involved changing secondary school start times

(Robinson, 2016). To address these challenges and to understand the impact of school-level decisions and policies that do not necessarily involve the introduction of a new intervention, the EEF in 2019 introduced a new funding stream titled "Researching school choices." The studies funded as part of this stream look at how the different choices schools make lead to different outcomes by examining natural variation in the system and using QEDs to estimate the impact of different approaches. As is the case with EEF trials, there is also a strong emphasis on best practice, research transparency, and lack of conflicts of interest when undertaking these studies.

Teacher choices. In 2018, the EEF undertook a review of its grant-making process with the aim of understanding how to make its projects timelier and more relevant for schools. The review identified that head teachers and teachers are keen for the EEF to answer research questions that can more directly feed into the existing teaching practice. As with the "School choices" strand of work, these questions are often not related to the impact of manualized programs, which require schools to purchase particular resources or training. Instead, they are about the everyday choices that teachers make when planning their lessons and supporting their pupils such as: Does phoning home improve student behavior? Does marking books lead to more learning than whole-class feedback? and What are the most effective ways to read with a class? The EEF has recently launched a pilot through which to explore innovative evaluation designs to investigate such questions, including approaches (e.g., within-participant designs and more proximal outcomes) that mean trials can run over shorter time frames and with smaller numbers of schools than in typical EEF trials.

Few EEF-funded trials have shown interventions to work better than standard practice

Lortie-Forgues and Inglis (2019) recently reanalyzed RCTs commissioned by the EEF and the NCEE to generate insights into the effectiveness and informativeness of the approach of using large-scale educational trials to generate evidence. They assessed the magnitude and precision of effects found in 141 RCTs involving 1,222,024 students, commissioned by the two organizations. They found that many of these RCTs have produced small effects (mean effect size .06 SDs), with wide confidence intervals (mean width .30 SDs). The authors concluded that many of these trials are uninformative, which provides a narrow view on the use of RCTs and the What Works movement, particularly given its specific focus on the precision of the headline finding (Nevill, 2019a). The quality of trials has progressively improved based on lessons from the EEF and NCEE experience. A fundamental premise of the What Works movement is to keep learning not just about what works but also about how to best research what works. A good example of this is that the effective sample size of EEF trials over time has risen over time (and almost doubled in 2014 when

compared to earlier trials), meaning that EEF trials have become progressively more informative (Sanders, 2019).

In addition, it is essential that the What Works movement is not only able to say what does work but also what does not, as resources spent on ineffective practices could be better used elsewhere. The message that few popular programs available to schools are better than what schools are already doing (business as usual) is useful. What is even more valuable is to reflect on why only a few programs generate positive effects and explore how best to refine the EEF's research questions and designs to be more fit for purpose. This exploration should involve developing a more in-depth understanding of business as usual, as the lack of intervention impacts may reflect high-quality teaching practice, and generating better data on costs. It is essential to investigate what works for whom, where, and under what circumstances, which presents additional methodological and logical challenges (e.g., the sample sizes required to test the intervention impacts on subgroups).

So far, the EEF has been mostly responsive to what is available in the education system when choosing which interventions to fund and evaluate. The "best picks" tend to be those that have a clearer underlying theory of change and are driven by some prior evidence. The EEF is increasingly funding more piloting and development work to improve intervention designs. In terms of its future direction, the EEF may be more likely to have a greater impact on pupil outcomes if it considers testing fewer, more intensive, and more theory-driven interventions, using richer, more "realist" RCT designs.

The EEF has also learned that the larger the scale of implementation, and size of the RCT, the harder it is to find educationally interesting effects compared to business as usual (Nevill, 2019b). There are examples of EEF trials that have shown positive effects in the efficacy testing stage (Hanley et al., 2015), which have not been replicated in the effectiveness testing stage (Kitmitto et al., 2018). There is often a direct tension between high-quality implementation and the need for greater statistical power. A large study sample may require intervention providers to scale up faster than appropriate, risking the scale-up mechanism. Smaller studies allow for more intensive support, monitoring, and engagement which keeps the processes of implementation more manageable. When increasing the scale of an intervention, providers will often need to recruit and train new staff and adapt training models (World Health Organization and ExpandNet, 2009). The EEF has recognized that scaling up is hard, whether this is within a large-scale study or post-experimentation. It now needs to rise to the challenge of investigating the key features that facilitate large-scale, successful, delivery of education interventions in different settings and contexts.

It is difficult to work out what works

Interpreting the results of RCTs can be a challenging task. A measure of what is "informative" as defined by the precision of the estimate misses many other important elements of security that

arguably tell us more about the reliability of the evaluation and its conclusions. It is for that reason that the EEF developed its padlock security rating designed to summarize, in a single scale, a number of possible sources of bias that could threaten the security of a result (The EEF, 2019c). While it might be controversial to summarize the quality of a headline finding in a single scale, it represents an important attempt to make evidence as accessible as possible to time-poor practitioners, so that they can better use it to inform their practice.

There are underlying problems with current analysis and reporting practices that exacerbate the challenges of communicating evidence and influence the interpretation of what works. The research community continues to practice the misinterpretation of p values and significance testing (Wasserstein et al., 2019), ranging from conclusions being based solely on whether an effect is found to be "statistically significant" and using arbitrary thresholds such as $p < .05$ to assign value to results, to concluding practical importance based on statistical significance or lack thereof. Furthermore, there remains a challenge regarding how we embrace uncertainty and report confidence intervals (Amrhein et al., 2019). An ongoing challenge for the EEF is to balance the adequate reporting of uncertainty with the need to communicate actionable results to end users.

Conclusion

The EEF has been influential in building capacity and capability to conduct high-quality education evaluations in the UK and is now recognized as a world leader in education evaluation (The Economist, 2018). It has strived for high standards of independent evaluation, pre-specification, and research transparency, analysis, and design. The lessons learned from EEF's journey can be valuable not only to other organizations but also governments and to those more widely trying to promote the What Works movement. For example, the EEF is now working collaboratively with other, newly formed, UK WWCs in areas including early intervention, crime and violence, employment, and social care, to share its learning, to develop best practice and ensure a consistent approach with respect to evidence generation. Furthermore, the EEF actively shares its lessons learned with new partner organizations globally. However, EEF's journey has not been without its challenges and no doubt more will emerge as the journey continues. The EEF is a learning organization and will strive to continue building on its experience to ensure it generates and shares the best possible evidence to improve teaching and learning and address educational disadvantage.

Authors' note

This article represents authors' views based on their experience of working at the EEF.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The Investing in Innovation Fund was established by the American Recovery and Reinvestment Act of 2009 and is administered by the U.S. Department for Education's Office of Innovation and Improvement, which asked the Institute for Education Sciences to provide support for conducting and supporting strong evaluations.
2. Impact evaluations include validation and scale-up evaluations, which are equivalent to Education Endowment Foundation's (EEF) efficacy and effectiveness evaluations. The EEF also commissions pilot evaluations which are equivalent to development grants in the U.S. but do not usually include an impact evaluation component.
3. Originally ISRCTN stood for "International Standard Randomised Controlled Trial Number"; however, over the years the scope of the registry has widened beyond randomized controlled trials to include any study designed to assess the impact of interventions in a human population.

Contributorship

Triin Edovald and Camilla Nevill conceived of the presented idea and were in charge of overall direction and planning. Both Edovald and Nevill contributed to the writing of the manuscript and provided critical feedback.

References

Allen, R., Jerrim, J., Parameshwaran, M., & Thompson, D. (2018). *Properties of commercial tests in the EEF database* (EEF Research Paper No. 001). https://educationendowmentfoundation.org.uk/public/files/Support/EEF_Research_Papers/Research_Paper_1_-_Properties_of_commercial_tests.pdf

Amrhein, V., Greenland, S., & McShane, B. (2019). Comment: Retire statistical significance. *Nature*, *305*, 567. https://www.nature.com/articles/d41586-019-00857-9

Anders, J., Brown, C., Ehren, M., Greany, T., & Nelson, R. (2017). *Evaluation of complex whole-school interventions: Methodological and practical considerations*. The Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Grantee_guide_and_EEF_policies/Evaluation/Setting_up_an_Evaluation/EEF_CWSI_RESOURCE_FINAL_25.10.17.pdf

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*(12), 2299–2306.

Boulay, B., Goodson, B., Olsen, R., McCormick, R., Darrow, C., Frye, M., Gan, K., Harvill, H., & Sarna, M. (2018). *The investing in innovation fund: Summary of 67 evaluations: Final report (NCEE 2018-4013)*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Burgess, S., Rawal, S., & Taylor, E. S. (2019). *Teacher peer observation and student test scores: Evidence from a field experiment in English secondary schools* (Ed Working Paper No. 19-139). http://www.edworkingpapers.com/sites/default/files/ai19-139.pdf

Cabinet Office. (2014). *What works? Evidence for decision makers* (Report). https://www.gov.uk/government/publications/what-works-evidence-for-decision-makers

Cabinet Office. (2019). *What works network*. (Original published 28 June 2013). https://www.gov.uk/guidance/what-works-network

Cabinet Office and HM Treasury. (2018). *The What Works Network: Five years on* (Report). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/677478/6.4154_What_works_report_Final.pdf

Cheung, A., & Slavin, R. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292.

Connolly, P., Biggart, A., Miller, S., O'Hare, L., & Thurston, A. (2017). *Using randomised controlled trials in education*. Sage.

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 Journal*, *337*, a1655. https://www.bmj.com/content/bmj/337/bmj.a1655.full.pdf

Cullinane, C. (2018). *Teacher Voice Omnibus Survey for Sutton Trust, March 2018*. https://www.suttontrust.com/research-paper/best-in-class-2018-research/

Dawson, A., Yeomans, E., & Rosa Brown, E. (2017). Methodological challenges from education RCTs: Reflections from England's education endowment foundation. *Educational Researcher*, *60*(3), 292–310.

Demack, S. (2019). *Does the classroom matter in the design of educational trials? A theoretical and empirical review* (EEF Research Paper No. 003). https://educationendowmentfoundation.org.uk/public/files/Publications/Does_the_classroom_level_matter.pdf

Department for Education. (2019a). *Schools, pupils and their characteristics: January 2019*. https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2019

Department for Education. (2019b). *School workforce in England: November 2018*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/811622/SWFC_MainText.pdf

Department for Education and The Rt Hon Michael Gove MP. (2010). *New endowment fund to turn around weakest schools and raise standards for disadvantaged pupils* (Press Release). https://www.gov.uk/government/news/new-endowment-fund-to-turn-around-weakest-schools-and-raise-standards-for-disadvantaged-pupils

Durlak, J., & DuPre, E. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*(3–4), 327–350.

Edovald, T., & Firpo, T. (2016). *Running randomised controlled trials in innovation, entrepreneurship and growth: An introductory guide*. Innovation Growth Lab at Nesta. https://www.innovationgrowthlab.org/guide-randomised-controlled-trials

FFT Education Datalab. (2018, January 23). *What is the national pupil database?* [Blog post]. https://fftedu cationdatalab.org.uk/2018/01/what-is-the-national-pupil-database/

Foliano, F., Rolfe, H., Buzzeo, J., Runge, J., & Wilkinson, D. (2019). *Changing mindsets: Effectiveness trial evaluation report*. The Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Changing_Mindsets.pdf

Franco, A., Malhotra, N., & Simonovits, G., (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505.

Ginsburg, A., & Smith, M.S. (2016). *Do randomized controlled trials meet the 'gold standard'? A study of the usefulness of RCTs in the What Works Clearinghouse*. American Enterprise Institute. https://pdfs.semanticscholar.org/3dee/5d1ab9a816c6299a7f2293c82542019eb41a.pdf

Goldacre, B., & Heneghan, C. (2014). Improving, and auditing, access to clinical trial results. *British Medical Journal*, *348*, g2130. https://www.bmj.com/content/348/bmj.g213

Hanley, P., Slavin, R., & Elliott, L. (2015). *Thinking, doing, talking science. Evaluation report and executive summary*. The Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Oxford_Science.pdf

Higgins, S., Katsipataki, M., Coleman, R., Henderson, P., Elliot Major, L., Coe, R., & Mason, D. (2015). *The Sutton Trust-Education Endowment Foundation teaching and learning toolkit. Manual*. The Education Endowment Foundation.

Higgins, S., Katsipataki, M., Kokotsaki, D., Coe, R., Elliot Major, L., & Coleman, R. (2013). *The Sutton Trust-Education Endowment Foundation teaching and learning toolkit: Technical appendices*. The Education Endowment Foundation.

Hodgen, J., Adkins, M., Ainsworth, S., & Evans, S. (2019). *Catch Up® numeracy. Evaluation report and executive summary*. The Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Catch_Up_Numeracy.pdf

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., & Michie, S. (2014). Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *British Medical Journal*, *348*, 1687–1699.

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016). *Implementation and process evaluation (IPE) for interventions in education settings: A synthesis of the literature*. The Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/IPE_Review_Final.pdf

Hutchinson, D., & Styles, B. (2010). *A guide to running randomised controlled trials for education researcher*. NFER.

Institute for Education Sciences. (2017). *Evaluation principles and practices*. National Centre for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/projects/pdf/IESEvaluationPrinciplesandPractices_011117.pdf

Kitmitto, S., González, R., Mezzanote, J., & Chen, Y. (2018). *Thinking, doing, talking science. Evaluation report and executive summary*. The Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/TDTS.pdf

Lord, P., Rabiasz, A., Roy, P., Harland, J., Styles, B., & Fowler, K. (2017). *Evidence-based literacy support: The 'literacy octopus' trial. Evaluation report and executive summary.* The Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Evidence_based_literacy_support_with_addendum.pdf

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*. http://eprints.whiterose.ac.uk/141754/3/LFI_2019_manuscript_supppdf

McNally, S., Ruiz-Valenzuela, J., & Rolfe, H. (2018). *ABRA: Online reading support. Evaluation report and executive summary*. The Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/ABRA_with_addendum.pdf (Original report published 2016; Addendum added 2018)

Montgomery, P., Grant, S., Mayo-Wilson, E., Macdonald, G., Michie, S., Hopewell, S., & Moher, D. (2018). Reporting randomised trials of social and psychological interventions: The CONSORTSPI 2018 extension. *Trials*, *19*(1), 1–14.

Morris, S., Edovald, T., Lloyd, C., & Kiss, Z. (2016). The importance of specifying and studying causal mechanisms in school-based randomised controlled trials: Lessons from two studies of cross-age peer tutoring. *Education Research and Evaluation*, *22*(7–8), 422–439.

Muralidharan, K., & Sundararaman, V. (2011).Teacher performance pay: Experimental evidence from India. *Journal of Political Economy*, *119*(1), 39–77.

Murphy, R., Weinhardt, F., Wyness, G., & Rolfe, H. (2017). *Lesson study. Evaluation report and executive summary*. The Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Lesson_Study.pdf

National Research Council. (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations*. The National Academies Press.

Nevill, C. (2016, April 5). *EEF blog: Do EEF trials meet the new gold standard?* [Blog post]. https://educationendowmentfoundation.org.uk/news/do-eef-trials-meet-the-new-gold-standard/

Nevill, C. (2019a, March 13). *EEF blog post: How do we make EEF trials as informative as possible?* [Blog post]. https://educationendowmentfoundation.org.uk/news/eef-blog-how-do-we-make-eef-trials-as-informative-as-possible/

Nevill, C. (2019b). *Reflections on eight years of commissioning education RCTs: What could we do different?* Presentation at the RSS event 'One Hundred Years of Education Trials: No Significant Difference?', London, UK.

Oakley, A. (2006). Resistances to 'new' technologies of evaluation: Education research in the UK as a case study. *Evidence and Policy*, *2*(1), 63–87.

Open Science Collaborative. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 943–951.

Quigley, A. (2019). *RCTs: What do they mean for teachers and school leaders?* Presentation at the RSS event 'One Hundred Years of Education Trials: No significant Difference?', London, UK.

Quigley, A., & Coleman, R. (2019). *Improving literacy in secondary schools. Guidance report*. The Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Publications/Literacy/EEF_KS3_KS4_LITERACY_GUIDANCE.pdf

Robinson, L. (2016). *Evaluation of the Teensleep programme*. *Evaluation protocol*. The Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/EEF_Project_Protocol_Teensleep_pilot.pdf

Roy, P., Styles, B., Walker, M., Bradshaw, S., Nelson, J., & Kettlewell, K. (2018). *Best practice in grouping students Intervention B: Mixed attainment grouping*. *Pilot report and executive summary*. The Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Intervention_B_-_Mixed_Attainment_Grouping.pdf

Sanders, M. (2019). *The challenges of being a trailblazer: Learning about learning* [Cabinet Office: What Works Blog post]. https://whatworks.blog.gov.uk/2019/03/08/the-challenges-of-being-a-trailblazer-learning-about-learning/

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin.

Shulz, K. F., Altman, D., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *British Medical Journal*, *340*, c332. https://www.bmj.com/content/340/bmj.c332

Styles, B., & Torgerson, C. (2019). *Trials, tribulations and centenary celebrations* (TES, No. 5365, pp. 16–17).

Sutherland, A., Prideaux, R., Bélanger, J., Broeks, M., Shenderovich, Y., & van der Staaij, S. (2017). *'Motivating teachers with incentivised pay and coaching' randomised control trial (ICR trial): Understanding factors influencing participant recruitment failure. Closure report*. The Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Motivating_teachers_with_incentivised_pay_and_coaching.pdf

The Economist. (2018, March 31). *England has become one of the world's biggest education laboratories*. https://www.economist.com/britain/2018/03/31/england-has-become-one-of-the-worlds-biggest-education-laboratories

The Education Endowment Foundation. (2012a). *Annual report 2011–2012*. https://educationendowmentfoundation.org.uk/public/files/Annual_Reports/EEF_Annual_Report_2011-12_-_interactive.pdf

The Education Endowment Foundation. (2012b). *EEF guidance on choosing and delivering attainment tests*. https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/EEF_testing_criteria_and_guidance_on_blinding.pdf

The Education Endowment Foundation. (2017a). *EEF standards for independent evaluation panel members.* https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/Evaluation_panel_standards.pdf

The Education Endowment Foundation. (2017b). *The attainment gap*. https://educationendowmentfoundation.org.uk/public/files/Annual_Reports/EEF_Attainment_Gap_Report_2018.pdf

The Education Endowment Foundation. (2018). *Statistical analysis guidance for EEF evaluations: March 2018.* https://educationendowmentfoundation.org.uk/public/files/Grantee_guide_and_EEF_policies/Evaluation/Writing_a_Protocol_or_SAP/EEF_statistical_analysis_guidance_2018.pdf

The Education Endowment Foundation. (2019a). *Guidance on recruitment and retention*. https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/Recruitment_and_retention_guidance_2019.pdf

The Education Endowment Foundation. (2019b). *Implementation and process evaluation guidance for EEF evaluators*. https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/IPE_guidance.pdf

The Education Endowment Foundation. (2019c). *Classification of the security of findings from EEF evaluations*. https://educationendowmentfoundation.org.uk/public/files/Evaluation/Carrying_out_a_Peer_Review/Classifying_the_security_of_EEF_findings_2019.pdf

Tirole, J., & Bénabou, R. (2006). Incentives and prosocial behavior. *American Economic Review*, *96*(5), 1652–1678.

Tymms, P., & Kasim, A. (2018). *Standard deviation as an outcome on interventions: A methodological investigation* (EEF Research Paper No. 002). https://educationendowmentfoundation.org.uk/public/files/Support/EEF_Research_Papers/Research_Paper_2_-_Standard_Deviation.pdf

U. S. Department of Education. (2009). *Race to the top program. Executive summary*. https://www2.ed.gov/programs/racetothetop/executive-summary.pdf

van Poortvliet, M., Axford, N., & Lloyd, J. (2018). *Working with parents to support children's learning. Guidance report*. The Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Publications/ParentalEngagement/EEF_Parental_Engagement_Guidance_Report.pdf

van Poortvliet, M., Clarke, A., & Gross, J. (2019). *Improving social and emotional learning in primary school.* https://educationendowmentfoundation.org.uk/public/files/Publications/SEL/EEF_Social_and_Emotional_Learning.pdf

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p<0.05". *The American Statistician*, *73*(S1), 1–19.

What Works Clearinghouse (WWC). (2017). *What Works Clearinghouse standards handbook version 4.0* (IES: WWC. 4.0). https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

White, H. (2019). The twenty-first century experimenting society: The four waves of the evidence revolution. *Palgrave Communications*, *5*(47), 2–7.

Wigelsworth, M. (2017). *SPECTRUM': Social, psychological, emotional, concepts of self, and resilience: Understanding and measurement*. Education Endowment Foundation.

World Health Organization and ExpandNet. (2009). *Practical guidance for scaling up health service innovations*. WHO Press.

Xiao, Z., Higgins, S., & Kasim, A. (2016). Same difference? Understanding variation in the estimation of effect sizes from educational trials. *International Journal of Educational Research*, *77*, 1–14.