



UNIVERSITY OF LEEDS

This is a repository copy of *Is the educational 'what works' agenda working? Critical methodological developments*.

White Rose Research Online URL for this paper:  
<http://eprints.whiterose.ac.uk/100081/>

Version: Accepted Version

---

**Article:**

Pampaka, M, Williams, J and Homer, MS [orcid.org/0000-0002-1161-5938](https://orcid.org/0000-0002-1161-5938) (2016) Is the educational 'what works' agenda working? Critical methodological developments. *International Journal of Research and Method in Education*, 39 (3). pp. 231-236. ISSN 1743-727X

<https://doi.org/10.1080/1743727X.2016.1170476>

---

**Reuse**

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.



[eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk)  
<https://eprints.whiterose.ac.uk/>

# Is the educational ‘what works’ agenda working? Critical methodological developments

## Editorial

### Motivation for the Special Issue

We began preparing this special issue with a widely accepted realisation: The search of ‘what works’ in educational contexts has been problematic in practice and in theory as well as in the interchange of the two via appropriate research methods and methodologies (e.g. Smeyers & Depaepe, 2010). The central idea from the wider ‘what works’ debate is about using evidence to make better decisions to improve public services. This gave rise to the so-called “evidence-based practice movement”, which is primarily linked to the use of Randomised Control Trials (RCTs) to ‘test’ interventions and measure their efficacy.

For education in the UK, the recent focus has been around “Improving education outcomes for school-aged children” led by Sutton Trust/Educational Endowment Foundation (EEF) and influenced by Ben Goldacre and the ‘nudge unit’<sup>1</sup>. Similarly, the US Department of Education<sup>2</sup>, based on the results from “high-quality research” (in programs, products, practices and policies) tries to answer the question “What works in education?” with the aim being to provide educators with the information they need to make “evidence-based decisions” via the ‘What works Clearinghouse’ (WWC). Thus, the topic of ‘what works’ is highly relevant and timely, especially in light of significant and dominant funding of interventions (e.g. the EEF initially had something like £50 million to invest within a relatively short time frame). It is hoped that some interventions will be shown to ‘work’, and this has been the focus for considerable RCT research: but thus far few significantly positive educational practices have emerged as working at large scale.

Such initiatives to align research with changes in practice are certainly not new as such; consider for example the idea of educational design research (Van den Akker, et al., 2006; Wittman, 1995) whose predecessors signal the need for research to influence change at least as far back as the 70s and 80s. They were also closely associated with the need for transformation of educational research and practice which led to a range of initiatives aimed at narrowing the gap between research, policy and practice, in UK and the call for causal analysis by means of experimental research in the USA.

Actually, a few of the papers in this special issue review the development and trajectories of the ‘what works’ agenda in education, and the closely related ‘evidence based education’ so we will skip a detailed (historic) overview in our editorial. Hanley, Cambers and Haslam (2016) for instance provide a very good ‘short history of

---

<sup>1</sup> <http://www.badscience.net/2012/06/heres-a-cabinet-office-paper-i-co-authored-about-randomised-trials-of-government-policies/>.

<http://www.theguardian.com/politics/2012/jun/20/test-policies-randomised-controlled-trials>

<sup>2</sup> “Evidence for What Works in Education”: <http://ies.ed.gov/ncee/wwc/default.aspx>

evidence-based education'. Spybrook, Shi and Kelcey (2016, [1 online](#)), in their introduction also refer to the 'first conference entitled "Randomised Controlled Trials in the Social Sciences" held at the University of York in 2006 which has since become an annual event that "brings in high-profile education researchers from various countries to discuss challenges and strategies for conducting high-quality RCTs in education".

Even though the call for RCTs remains dominant for 'evidence-based' decisions, there have been various opponents of this idea of such an evidence-based education, e.g. Biesta, (2007): see also Hammersley's (2005) critique of Chalmers (2003) initiative, and Thomas<sup>3</sup> (2012) urge to 'forge a new science of education'. Others have also prompted alternative approaches (e.g. Smithers, 1993). Furthermore, there has also been an ongoing discussion about what 'evidence' should entail, and consequently how to establish a balance, integration or synthesis (as put by Peterson in one of the contributions in this volume) between RCTs and other (qualitative and quantitative) approaches.

There is thus still an unresolved debate in educational research, involving the 'what works' agenda. This covers a spectrum of topics including performativity, effectiveness, equality, equity, bridging gaps (social class, gender, ethnicity, etc), assessment, improvement, and causality, and the best methods for investigating these issues. A key aspect of this is the need on all sides to ensure that research is both rigorous and relevant: Can this be achieved with current methods? Can we achieve both rigour in methods and also relevance to current educational concerns and decision-making? But maybe we need to take even a step back: have we even come to a consensus as a community about 'rigour' and 'relevance'. It was our intention with this special issue to provide the platform for sharing new perspectives and bringing together new methods/methodologies that can best contribute to this agenda/debate and thus can help towards 'working solutions' in the educational research arena, whether they essentially support the orthodoxy of 'what works' or raise a challenge to it.

## **Contributions for Part 1**

In response to the call for papers we have received contributions in a range of themes, which partly covered and went beyond our initial list. As expected, the dominant theme is still impact evaluations and educational RCTs, along with other commonly used methods in evidence-based research and practice, such as systematic reviews and meta-analysis. Other contributions cover measurement issues, single case studies and longitudinal designs, dealing with missing data and imputation techniques. There have also been contributions related to effective communication and dissemination of the relevant evidence/research findings in order to reach maximum impact with the relevant stakeholders involved. Given the number of papers that survived the reviewing process we have been allowed to present these in a two part special issue. A thematic split brings together, in this first part, 7 papers which deal with common practice in

---

<sup>3</sup> See also <http://www.timeshighereducation.co.uk/comment/opinion/no-one-can-control-for-a-sense-of-when-4-3-3-might-turn-the-game/2001371.article> ("Randomised controlled trials can't solve social policy conundrums any more than they can fit football to a formula")

educational interventions and communicating findings with the involved/ relevant stakeholders. We open with three papers that discuss orthodox 'what works' approaches to current evidence based educational initiatives.

Steve Higgins and Maria Katsipataki set up the scene of the current orthodoxy for evaluating interventions with the dominant UK funder: the Sutton Trust- Education Endowment Foundation (EEF) and their guidance through the Teaching and Learning 'Toolkit'. Their paper provides a review of the strengths and limitations of the comparative use of meta-analysis findings, utilising thus another fundamental approach within such evidence driven research. As they state meta-analysis as 'a method of combining the findings of similar studies to provide a quantitative synthesis' is in general challenging, but even more so in educational research because of the complexity in regards to scope, scale and diversity. In their paper they embark to add on the debate about the use of meta-analysis in educational research using examples from the EEF toolkit in terms of communication of findings and drawing inferences. In doing so they introduce the 'super-synthesis' approach, a comparative meta-analysis looking at different meta-analytic studies with a common populations: the debates, issues and challenges around the use of such approach are described with the authors concluding that 'so long as you are aware of the limits of inferences drawn, then the approach has value' (p. 5). The paper also presents alternative ways of interpreting security ratings, cost- estimates effect sizes for accessibility of the findings in the relevant stakeholders. For example, they explain how they convert in the 'toolkit' effect sizes into a single scale of school progress in months. Their concluding suggestion is to consider evidence for 'what has worked' and not only 'what works' for more informed development of research-based practice.

The next two papers have technical interest for those involved in educational evaluations involving RCTs and aspire to advance existing methods. Studies like this that question the reliability and validity of self-declared 'gold standard' approaches from an 'internal' point of view should be welcomed and encouraged, in our view. In both cases the papers show how technical improvements might help make the approach 'work better' in practice, and so help justify the effort put into this orthodox approach in future (this is important given the current malaise that seems to be settling on the approach due to its failure to identify interventions that work at large scale). In each case we might raise the question, however, will such improvements as are suggested really make the necessary difference?

Jessaca Spybrook, Ran Shi and Benjamin Kelcey report on recent (last decade) progress with RCTs: their focus is on the design of cluster randomized trials (CRT) which are becoming a standard practice, as typically interventions are implemented at school or classroom level, rather than individual pupil level. They are particularly interested in the capacity of such studies to 'provide rigorous evidence of whether or not a programme is effective' (p. 2). Rigorous evidence is viewed in their paper in respect to the statistical power with which the study is designed 'for an effect of a given magnitude'. In order to do so the paper clarifies, even for the unfamiliar reader, the concepts of power, effect size, and 'minimum detectable effect size' (MDES). MDES, a

measure of study precision is actually focus of debates recently as its mainly governed by rules of thumps. In order to map recent progress in this regard Spybrook et al compare the design and precision of studies funded by NCER in the early years (2002-2004, cohort 1) to those in the recent years (2011-2013, cohort 2). Their comparisons of the total number of clusters randomised and MDES aim to examine whether there has been a shift in the quality of the design of CRTs and are based on 38 studies (16 from cohort 1 and 22 from cohort 2). A valuable part of this paper then is the inclusion of the formulas for the calculation of minimum detectable effect size for 2 and 3-level CRTs which fill in a gap in current methodological literature and we expect to help evaluators in powering their studies. As the authors conclude, the results of this study 'provide evidence of an increase in the precision of CRTs funded by NCER'. They then provide some potential explanations for this change in precision over time, including the increase in knowledge base and the increase in professional development opportunities around the design of CRTs.

In a similar vein, Jonathan Schweig and John Pane focus on Partially-Nested RCTs which address similar concerns to those of Spybrook et al: they note that even when individuals are randomised, most often both experimental and control students share common learning experiences as they are grouped in classrooms and schools. They then focus on the less common, partially clustered design which sees one clustered and one unclustered experimental arm, in light of real-world complications with RCTs. The complication they choose to address regards treatment noncompliance and the lack of existing methodological literature addressing this issue within partially nested designs. As they state 'noncompliance is an issue of practical importance in educational research' and in its presence all the statistical methods proposed for RCT designs, within the multilevel model framework, 'no longer estimate the effect of treatment, because they do not model the actual treatment received by each individual' (p.?). Instead such models usually offer intent-to-treat (ITT) estimates which is of great importance for policy and practice 'as policymakers and administrators often have control only over the availability of an intervention, and not its uptake'. After outlining analytical challenges presented by noncompliance in Partially Nested RCT designs, the paper goes on to offer two possible resolutions, with amendments to existing models. In order to verify the appropriateness of these resolutions they then proceed with a simulation study based on data taken from a large scale RCT. This on its own is a useful demonstration for educational researchers facing similar problems. Their results provide evidence that 'clustering and noncompliance can have substantial impacts on statistical inferences about ITT effects'. They conclude with some potentially useful resolutions for analysts who face such complications in their designs. We might further question what constitutes 'non-compliance': if indeed even a proportion of those technically 'compliant' practitioners are not effectively willing participants these results might suggest another reason why large scale implementations often fail in their promise.

Both the following two papers then question the 'superiority' of RCTs as the norm for 'What works' and suggest alternative approaches that integrate other methods. A reassessment of RCTs for evaluating educational interventions is offered by Pam

Hanley, Bette Chamber and Jonathan Haslam. In this contrasting/different perspective, the authors problematize the use of the term 'gold standard' to describe RCTs because it implies inferiority of other methods. As they argue their usefulness 'can be greatly enhanced when used in conjunction with implementation-specific measures' (p.1). In the two evaluation case studies they then present they show how a process evaluation is integrated within the overall evaluations to provide evidence from classroom practice along with feedback from teachers and students. Those involved in evaluations of programmes involving school change of practice and the complexities inherent in such efforts would concur with the authors view that such approach allows for 'more holistic and richly interpretative pieces of research' (p.1). A key idea from the case studies is that it takes time for new practices to develop and become effective, and one might thus expect in the first year of implementation of a substantially new practice (like cooperative learning): we might suggest even that classroom practice to may even 'get worse' from certain points of view. Teachers need time to make mistakes, learn from them, and develop accordingly in any kind of significantly new practice. Measuring and comparing only 'hard learning outcomes' may also be misleading. Therefore Hanley et al (2016) argue against privileging particular paradigms in designing effective evaluations: 'Perhaps it is time for the bluntness of the "what works?" agenda to evolve into one that also establishes who it works for, through what means, and in what circumstances' (p. 10).

Amelia Peterson, in the next paper in this volume, approaches this issue from the lens of integration of experimental and improvement science. Starting from the contention that current approaches to 'What works' mainly involving RCTs limits the potential for impact at scale, she then proposes alleviating some of the problems of the so-called 'Black-box RCTs' initiated by Scriven (1994). The term refers to the lack of information provided by trials as to how, why and under what conditions the programme works. Such an approach entails various limitations including poor fit with teacher-led interventions, weak implementation and scale, and low adaptation as detailed by Peterson. In order to alleviate these limitations she then proposes 'What works 2.0' which combines the core elements of experimental and improvement science into a strategy to raise educational achievement with the support of evidence from randomized experiments. 'Central to this combined effort is a focus on identifying and testing mechanisms for improving teaching and learning, as applications of principles from the learning sciences' (Peterson, 2016, p.1 citing Bransford et al. 2000 and Della Chiesa et al. 2007). (p.1)

Similar ideas are shared by Sardar Anwaruddin who approached this from a different stance in the next paper: while addressing the crisis of representation (i.e. the challenges of representing the lived experiences of research participants, Denzin and Lincoln, 2005) in educational research, he reflects on the effects of the 'what works' on educational research. He does so by presenting the insights of teachers, through their participation in an online (wiki) discussion and via individual interviews, with research in their reading, interpretation and adoption of these from a published research article. The analysis of the data along with the theorizing of the results within recent philosophical traditions bring to light issues surrounding the communication and dissemination of

research. The author situates this within the concepts of 'knowledge mobilization' and Hegel's (1977) 'recognition' and problematizes the dominant method of knowledge management within 'what works' paradigms, that is 'the transfer of research-based knowledge to practice contexts such as schools' (p. 11). This calls for such agendas to rethink this transfer model and, as he concludes, to invite the teachers' perspective into attempts to understand how the crisis of representation might be overcome.

A closely related issue is raised in the final paper by Chris Green, Celia Taylor, Sharon Buckley and Sarah Hean. In their paper 'Beyond synthesis: augmenting systematic review procedures with practical principles to optimise impact and uptake in educational policy and practice' as the title suggests do a very good job to bring the representation of research participants, and other relevant stakeholders in central positions in the process of developing evidence-based studies and reviews. Towards this direction they introduce the 'Beyond Synthesis Impact Chain' (BSIC), a framework for producing and disseminating conclusions meaningful for practitioners and policy-makers. This framework brings at the foreground the issue of impact of research and proposes the involvement of the relevant stakeholders in systematic reviews and synthesis of existing evidence. The examples they draw on to contextualise their proposed framework come from the health profession education literature, however the principles and strategies described are applicable to any educational problem under review.

### **Getting these together and next steps**

The papers included in this issue, as introduced, focus primarily on the central mainstream foci of a 'what works' agenda including RCTs, and alternative integrated approaches to evaluations. The topics covered ranged from language teaching (Anwaruddin), health profession education (Green et al), science and maths teaching (Hanley et al) and voluntary summer learning programs (Schweig and Pane).

This issue also covers a range of methodological approaches including systematic reviews and meta-analysis (Green et al, Higgins and Katsipataki), advancements of methods for dealing with the complexity involved in interventions by either improved models (Spybrook et al, Schweig et al, this volume) and/or integrated approaches (Hanley et al, Peterson, this volume).

The first (Higgins and Katsipataki) and the last two papers (Anwaruddin; Green et al) are also related to dissemination of findings with the relevant stakeholders addressing directly the impact agenda of educational research which is becoming prominent in the various funding agendas (e.g. Economic and Social Research Council, ESRC) as well as academic/research evaluation reformed agendas (e.g. Research Excellence Framework).

As for our initial question and aim to explore whether the 'what works' agenda is working, we might be half way there: Papers in this issue addressed and questioned/proposed approaches for ensuring we cover not just 'what works?', but also "why something works", 'what has worked? Where? How? For whom?' etc. In this view

practice needs a much wider knowledge base than policy (which seeks simplicity and therefore puts a premium on supporting specific programmes). There seems to be, thus a paradigmatic split depending on the question asked. A key philosophic difference between these two approaches lies in answering a fundamental question about 'agency': the former approach assumes all agency is in the hands of policy and programme, and they want the intervention to be teacher-proof. The latter want to allow for variation or even mediation by teachers and other local factors, giving more agency in the whole process to schools, and teachers. Arguably, however, neither of them see the learners as having agency: this has not been covered yet in this emerging debate. In previous work in this lines the emphasis of current testing/assessment practices at all, has been questioned proposing the measurement of 'alternative learning outcomes' (Williams & Ryan, 2013; Pampaka et al. 2013). In particular, attitudes, dispositions, and aspirations might need to be measured as well as attainment if we are to capture the complexities of teaching-learning relationships.

These are issues we hope to pursue more vigorously in future issues concerning 'what works' (Part 2), from more robustly critical perspectives.

## References

Anwaruddin, S. (2016) Language teachers' responses to educational research: addressing the 'crisis' of representation, *International Journal of Research and Method in Education*, 39, 3, [pages to be added once SI is typeset]

Biesta, G. (2007). Why "what works" won't work: evidence-based practice and the democratic deficit in educational research. *Educational Theory*, 57(1), 1-22.

Chalmers, I. (2003). Trying to do more good than harm in policy and practice: the role of rigorous, transparent, up-to-date evaluations. *Annals of the American Academy of Political and Social Science*, 589, 22-40.

Denzin, N.K. & Lincoln, Y.S. (eds) (2005). *The Sage Handbook of Qualitative Research*. London: Sage.

Green, C., Taylor, C., Buckley, S. and Hean, S. (2016) Beyond synthesis: Augmenting systematic review procedures with practical principles to optimise impact and uptake in educational policy and practice, *International Journal of Research and Method in Education*, 39, 3, [pages to be added once SI is typeset]

Hammersley, M. (2005). Is the evidence-based practice movement doing more good than harm? Reflections on Iain Chalmer's case for research-based policy making and practice. *Evidence & Policy*, 1(1), 85-100.

Hanley, P., Cambers, B. and Haslam, J. (2016) Progress in the Past Decade: An Examination of the Precision of Cluster Randomized Trials Funded by the U.S. Institute

of Education Sciences, International Journal of Research and Method in Education, 39, 3, [pages to be added once SI is typeset]

Hegel, G. W. F. (1977). Phenomenology of Spirit. Oxford: Clarendon Press.

Higgins, S. & Katsipataki (2016) Communicating comparative findings from meta-analysis in educational research: some examples and suggestions, International Journal of Research and Method in Education, 39, 3, [pages to be added once SI is typeset]

Pampaka, M., Williams, J. S., Hutchenson, G., Black, L., Davis , P., Hernandez-Martines, P., & Wake, G. (2013). Measuring Alternative Learning Outcomes: Dispositions to Study in Higher Education. Journal of Applied Measurement, 14(2), 197-218.

Peterson, A. (2016) Getting 'What Works' working: building blocks for the integration of experimental and improvement science, International Journal of Research and Method in Education, 39, 3, [pages to be added once SI is typeset]

Schweig, J. and Pane, J. (2016) Intention-to-Treat Analysis in Partially-Nested Randomized Controlled Trials with Real-World Complexity, International Journal of Research and Method in Education, 39, 3, [pages to be added once SI is typeset]

Scriven, M (1994). The fine line between Evaluation and Explanation. American Journal of Evaluation, 15(1). 75-77.

Smithers, A. (1993). All Our Futures: Britain's Education Revolution. A dispatches report on Education.

Smeyers, P. & Depaepe, M. (Eds) (2010). Educational Research: Why 'What Works' Doesn't Work. Springer.

Spybrook, J., Shi, R. and Kelcey, B. (2016) Progress in the Past Decade: An Examination of the Precision of Cluster Randomized Trials Funded by the U.S. Institute of Education Sciences International Journal of Research and Method in Education, 39, 3, [pages to be added once SI is typeset]

Thomas, G. (2012) Changing our landscape of inquiry for a new science of education. Harvard Educational Review, 82, 1, 26-51.

Van den Akker, J. Gravemeijer, K., McKenney, S. & Nieveen, N (Eds). 2006. Educational Design Research. Routledge.

Williams, J. & Ryan, J. (2013). Research, policy, and professional development." In What works in education?, ed. V. Farnsworth & Y. Solomon, London: Routledge.

Wittman, E.Ch. (1995). Mathematics Education as a "Design Science". *Educational Studies in Mathematics*, 29, 355-374.