NFER | Educational Research

Routledge
Taylor & Francis Group

Educational Research

Special Issue: Randomised controlled trials (RCTs) in education research – methodological debates, questions, challenges
Guest Editors: Ben Styles and Carole Torgerson

ISSN: 0013-1881 (Print) 1469-5847 (Online) Journal homepage: https://www.tandfonline.com/loi/rere20

The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016

Paul Connolly, Ciara Keenan & Karolina Urbanska

To cite this article: Paul Connolly, Ciara Keenan & Karolina Urbanska (2018) The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016, Educational Research, 60:3, 276-291, DOI: 10.1080/00131881.2018.1493353

To link to this article: https://doi.org/10.1080/00131881.2018.1493353

Published online: 09 Jul 2018.

EDUCATIONAL RESEARCH
2018, VOL. 60, NO. 3, 276–291
https://doi.org/10.1080/00131881.2018.1493353

The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016

Paul Connolly [a], Ciara Keenan [a] and Karolina Urbanska [b]

[a]Centre for Evidence and Social Innovation, Queen's University Belfast, Belfast, Northern Ireland; [b]School of Psychology, Queen's University Belfast and LAPSCO, Université Clermont Auvergne, Clermont-Ferrand, France

ABSTRACT

Background: The use of randomised controlled trials (RCTs) in education has increased significantly over the last 15 years. However, their use has also been subject to sustained and rather trenchant criticism from significant sections of the education research community. Key criticisms have included the claims that: it is not possible to undertake RCTs in education; RCTs are blunt research designs that ignore context and experience; RCTs tend to generate simplistic universal laws of 'cause and effect'; and that they are inherently descriptive and contribute little to theory.

Purpose: This article seeks to assess the above four criticisms of RCTs by considering the actual evidence in relation to the use of RCTs in education in practice.

Design and methods: The article is based upon a systematic review that has sought to identify and describe all RCTs conducted in educational settings and including a focus on educational outcomes between 1980 and 2016. The search is limited to articles and reports published in English.

Results: The systematic review found a total of 1017 unique RCTs that have been completed and reported between 1980 and 2016. Just over three quarters of these have been produced over the last 10 years, reflecting the significant increase in the use of RCTs in recent years. Overall, just over half of all RCTs identified were conducted in North America and a little under a third in Europe. The RCTs cover a wide range of educational settings and focus on an equally wide range of educational interventions and outcomes. The findings not only disprove the claim that it is not possible to do RCTs in education but also provide some supporting evidence to challenge the other three key criticisms outlined earlier.

Conclusions: While providing evidence to counter the four criticisms outlined earlier, the article suggests that there remains significant progress to be made. The article concludes by outlining some key challenges for researchers undertaking RCTs in education.

ARTICLE HISTORY
Received 11 December 2017
Accepted 22 June 2018

KEYWORDS
Systematic reviews; RCT; education; educational settings; context; theory

Introduction

Since the late 1990s there has been an increasing shift towards the notion of evidence-based practice in education (Thomas and Pring 2004; Hammersley 2007; Bridges,

CONTACT Paul Connolly paul.connolly@qub.ac.uk

Smeyers, and Smith 2009). A significant element of this has been concerned with research that has sought to identify and provide robust evidence of 'what works' in relation to educational programmes and interventions. Within this, randomised controlled trials (RCTs) have played a central role in seeking to determine whether an intervention is having a discernible and measurable effect on students' learning and development (Torgerson and Torgerson 2001, 2008; Connolly et al. 2017). The rationale underpinning the RCT is deceptively simple: it seeks to measure the progress of students participating in an educational intervention against that of a control group of equivalent students who are, most typically, continuing as normal. The key test is whether the progress made by those in the intervention group exceeds those in the control group, on average. If this is the case, the logic of the RCT suggests that this difference in progress is likely to be due to the effects of the intervention. Such a claim can be made only if the intervention and control groups are equivalent, and the creation of matched groups is achieved in an RCT by students being randomly allocated to both. If random allocation is undertaken properly and if the size of the two groups is sufficiently large, then the only systematic difference between the two groups of students is that one has received the intervention whilst the other has not. The process of randomisation has ensured that all of the other potential factors that may influence a students' progression are likely to be evenly distributed across the two groups. Differences can arise by chance but the statistical tests used to analyse such data take this into account.

However, whilst such a research design is simple, it has attracted trenchant criticism from many within the education research community (Connolly et al. 2017; Gorard, See, and Siddiqui 2017). Alongside claiming that it is simply not possible to conduct RCTs in education, critics suggest that RCTs ignore context and experience, that they tend to generate simplistic universal laws of 'cause and effect' and that they are inherently descriptive and contribute little to theory generation or theory building. The purpose of this article is to assess these criticisms by considering the actual evidence of how RCTs have been undertaken in practice in education. More specifically, the article is based upon a systematic review of all RCTs that have been undertaken and published in education between 1980 and 2016. By documenting and analysing how RCTs have been used and reported over this period, the article seeks to distinguish between some of the rhetoric and reality associated with these on-going debates. In this sense the article is concerned with the trials of evidence-based practice in education in two respects: the way that evidence-based practice has been put on trial and the accuracy of the charges levelled against it, and with better understanding the nature and focus of the actual trials themselves.

Background

The nature and tone of the criticisms levelled at the use of RCTs in education can be illustrated through the use of a few quotations. Perhaps the most notable critique of RCTs in education, within the United Kingdom at least, is that contained in the seminal methodology textbook by Cohen, Manion, and Morrison (2011), now in its seventh edition. *Research Methods in Education* teaches students of educational research the following about RCTs:

> This model [the RCT], premised on notions of isolation and control of variables in order to establish causality, may be appropriate for a laboratory, though whether, in fact, a social situation either ever could become the antiseptic, artificial world of the laboratory or should become such a world is both an empirical and a moral question respectively. Further, the ethical dilemmas of treating humans as manipulable, controllable and inanimate are considerable [. . .] Randomised controlled trials belong to a discredited view of science as positivism (p. 314).

Such criticism of RCTs is also evident in the arguments made by Hodkinson and Smith (2004: p. 151), who also suggest that the purpose of trials is to establish universal and replicable laws:

> Here, all variables are held constant except the one under investigation. Ideally, this one variable is deliberately changed, in two exactly parallel situations as, for example, when a new medical drug is tested against a placebo. If a difference is noted, rigorous tests are conducted to minimize the chances that it is coincidental. Laboratory experiments are repeated, to ensure the results always turn out in the same ways. [. . .] This is the view of research that lies at the heart of the evidence-informed movement.

Morrison (2001: p. 72–4) takes this criticism further by suggesting that RCTs promote a simplistic, decontextualised and atheoretical picture of the social world:

> Chaos and complexity theories here are important, for they argue against the linear, deterministic, patterned, universalisable, stable, atomised, objective, controlled, closed systems of law-like behaviour which may be operating in the world of medicine and the laboratory but which do not operate in the social world of education [. . .]

> The importance of context is undeniable, yet where is this taken into account in the RCT? The RCT actively builds out and excludes key elements of context, as that could 'contaminate' the experiment, yet it could be these very factors that are important.

These sentiments are also conveyed by Elliott (2004: 175–6), who develops this notion of RCTs failing to engage with context by contrasting RCTs with case studies:

> The primary role of *educational* research, when understood as research directed towards the improvement of *educational* practice, is not to discover contingent connections between a set of classroom activities and pre-standardised learning outputs, but to investigate the conditions for realising a coherent *educational* process in particular practical contexts. Both the indeterminate nature of educational values and principles, and the context-dependent nature of judgements about which concrete methods and procedures are consistent with them, suggest that *educational* research takes the form of case studies rather than randomised controlled trials. The latter, via a process of statistical aggregation, abstract practices and their outcomes from the contexts in which they are situated. Case studies entail close collaboration between external researchers and teachers on 'the inside' of an educational practice.

The above quotations are typical of the critical discourse that has surrounded the use of RCTs in education; a discourse that has largely remained unchanged for the past decade (Connolly et al. 2017). The underlying criticisms running through these quotations can be distilled down to four key charges: (1) that it is just not possible, on a practical level, to undertake RCTs in education; (2) that RCTs ignore context and experience; (3) that RCTs seek to generate universal laws of 'cause and effect'; and (4) that RCTs are inherently descriptive and contribute little to theory.

Through a systematic review of all RCTs conducted in education from 1980 to 2016, this article seeks to assess these four criticisms against the actual evidence. Following an outline of the methods used for this systematic review, the article sets out the key findings and then returns to these four criticisms and assesses them in the light of these findings. In doing this, the article also considers the implications of these findings for the future use of RCTs in education.

Methods

The methods used for this systematic review adhere to the guidelines recommended by the Centre for Reviews and Dissemination (2009), the Campbell Collaboration (2016) and the Cochrane Collaboration (Higgins and Green 2011). An initial scoping search of existing systematic reviews of RCTs in education was employed using the following databases: the Cochrane Library of Systematic Reviews; the Database of Abstracts of Systematic Reviews; and the Campbell Library. No existing, or planned, review was found that summarises RCTs in education across multiple educational institutions and with at least one educational outcome.

Inclusion criteria

The criteria for the inclusion of studies were developed using the PICOS method (i.e. Population, Intervention, Comparison, Outcomes and Study design), adapted from the original PICO method (Richardson et al. 1995) that enables researchers to itemise broad research questions into both searchable keywords and clear inclusion criteria. The criteria used for this review are summarised in Table 1. As shown, strict criteria were applied to ensure that all RCTs identified for inclusion in this systematic review are clearly and unambiguously regarded as education research.

Population: For the purposes of this systematic review, the population of interest is anyone who is the recipient of teaching, instruction or training. This includes anyone acquiring new knowledge and/or skills (i.e. students, teachers, doctors, medical students). Interventions that involve targeting particular sub-groups of learners are also included (i.e. those that only target children who are poor readers or only those with behavioural difficulties). Within this, teaching, instruction and/or training must either

Table 1. Summary of inclusion criteria used for the systematic review*.

Criterion	Description
Population	Any recipient of teaching, instruction or training.
Intervention	Eligible interventions must include an educational/learning component and be delivered by an educational institution or delivered explicitly through, and with the cooperation of, the educational institution.
Comparison	Studies must include a comparison or control group. Control groups may include placebo (no or sham treatment), treatment as usual and/or wait list.
Outcomes	Studies must include at least one educational outcome, defined broadly as relating to the acquisition of knowledge and/or skills.
Study design	Only studies involving the random allocation of subjects (either individually or as groups) to a control group and at least one intervention group to be included. These include: Randomised Controlled Trials; Randomised Cross-over Trials; and Cluster Randomised Controlled Trials.

*Authors original, unpublished table.

take place within an educational institution or be provided by an educational institution. Studies where the student is not the focus, such as parents being trained at home to deal with their child's behaviour, are not included.

Interventions: Eligible interventions must include an educational/learning component and be delivered by the educational institution (i.e. involving the teachers) or delivered explicitly through, and with the cooperation of, the educational institution. This includes interventions that use external facilitators coming into schools: while delivered by someone external, they are still delivered in association with the class teachers and schools who plan them into the school day.

One key distinction is that the intervention needs to be seen as part of the general/core business of that institution rather than just using the institution as a convenient method for reaching students. Thus, for college/university students, for example, eligible interventions would be restricted to those that relate directly to their learning within the institution. These could involve interventions that focus on enhancing students' knowledge and skills in relation to a particular element of their course and/or those that focus on broader study skills that seek to enhance their performance on that course. Interventions that simply used the institution to access students to test out non-educational interventions (i.e. smoking cessation programmes; alcohol reduction programmes; date violence reduction programmes; nutrition or dieting programmes) would not be eligible.

Eligible interventions can include those delivered off-site in the case of afterschool clubs or universities with students out on placement (most commonly medical students in hospitals). In such cases, the interventions must still satisfy the requirement that they are organised by, and directly related to, an educational institution and focused on enhancing the effectiveness of the students' learning. Learning environments which have not been organised by an educational institution – such as Saturday schools, activities run by other groups or societies (e.g. guides/scouts/cubs), supplementary schools and community schemes (i.e. diet clubs, summer schemes and football courses) – are all excluded.

Comparison: Only studies that include a comparison or control group are eligible for inclusion in this systematic review. Eligible studies include those with control groups that consist of a 'treatment as usual' condition or that may possibly use a placebo. Studies can include those that use wait list control groups (i.e. groups that continue as normal during the period of the intervention but then who receive the intervention after the trial has been completed), so long as the control group only receives the intervention once final post-testing has been completed.

Outcomes: For a study to be eligible, it must include a focus on evaluating the effects of an intervention on at least one educational outcome. Educational outcomes include any that relate to the acquisition of knowledge and/or skills (i.e. common curriculum subjects; broader cognitive skills; socio-emotional learning; vocational training and skills). School-based physical activity programmes that aim to change behaviour by increasing exercise with a goal to reduce obesity or BMI would not be included as there is no learning or development for the children. However, physical/movement programmes *would* count in early years settings if they are aimed at supporting children's development (i.e. where the outcomes are improved fine or gross motor skills development).

Study design: Included studies must involve the random allocation of subjects (either individually or as groups) to a control group and at least one intervention group. All studies without a control group or those that are quasi-experimental or do not randomly assign participants are excluded.

Search strategy

Searches of the literature were conducted to identify all RCTs conducted in education that met the above inclusion criteria. There was no limitation to publication type but the search excluded those papers published prior to January 1980 and unavailable in the English language. The search strategy used in this review is based on the Pearl Harvesting method developed by Sandieson (2006, 2017). The second author (Ciara Keenan) has been trained in this method directly from the developer and was responsible for all searches carried out through this review. The search strategy was created by developing free text terms based on relevant keywords related to educational outcomes, educational institutions and randomised trials. The search filters created from this method were then used directly in the command lines of databases and combined using Boolean operators. Each search filter was then adapted specifically for use in various databases. An example of the search strategy used within the ERIC (ProQuest) is presented in Table 2.

The search strategy was undertaken in various electronic databases and through grey literature sources covering outputs published from 1 January 1980 to 31 December 2016. Electronic databases searched were: CENTRAL (Cochrane Central Register of Controlled Trials – Wiley); ERIC (Educational Resources Information Center – ProQuest); BEI (British Education Index – EBSCOhost); PsycINFO (Ovid); and IBSS (International Bibliography of the Social Sciences – ProQuest).

The exclusion of grey literature in a systematic review of interventions is not only a threat to validity but also increases the risk of publication bias affecting results. Publication bias most simply refers to the likelihood that those studies with negative effects or non-statistically significant findings will not be published (Rosenthal 1979). To counteract the negative effects of publication bias, various grey literature sources were included in this review of interventions, including a database search of dissertations and theses (Global version – ProQuest), EEF (Education Endowment Foundation), WWC (What Works Clearinghouse) and hand-searching of relevant systematic reviews.

Study selection

Titles and abstracts of studies returned by the search strategy implemented through database searching were imported to a bibliographic reference manager (Refworks), where duplications of studies were removed. These abstracts were then uploaded to the web-based screening tool, *Abstrackr*, and all authors were invited to screen the abstracts independently. Reviewers made decisions to include, query or exclude an abstract. All decisions were then exported to Excel (2016) to check for consensus and disagreements. Those studies which both authors independently agreed to include moved forward for full text screening, and those studies which both authors independently agreed to exclude were removed from the library. For those studies which authors had queried or disagreed upon, consensus was reached through discussion.

Table 2. Example of search strategy used with the ERIC database*.

ti(Child* OR youth* OR pupil* OR 'young people' OR 'young persons' OR student* OR boy* OR girl* OR adolescen* OR teen* OR apprentic* OR tutor* OR mentor* OR teacher* OR traine*)
AND ti(educat* OR teach* OR councel* OR learn* OR impact* OR curricul* OR train* OR therap* OR instruct* OR achieve* OR program* OR treatment* OR evaluat* OR intervention* OR model* OR practice* OR vocation* OR academic* OR school* OR 'after-school' OR 'after school' OR classroom* OR class OR preschool* OR 'pre school*' OR kindergarten*OR Nurser* OR "early childhood education* " OR 'primary education' OR Kindergarten OR elementary OR "Primary class* " OR "Primary school* " OR "reception class* " OR Post-primary OR "Secondary school* " OR "Junior high " OR 'Middle school' OR 'elementary education' OR 'elementary school' OR 'high school' OR college OR universit*)
AND ti(trial* OR RCT* OR 'evaluat* study' OR 'evaluat* research' OR 'effectiv* study' OR 'effectiv* research' OR 'treatment effect*' OR 'control* study' OR 'control* studies' OR 'control* design*' OR 'control* trial*' OR 'control* group*' OR 'control group design' OR 'trial registration' OR 'quantitative research' OR CONSORT OR GRADE OR 'untrained control group*' OR 'control class*' OR 'comparison group*' OR 'positive-control study' OR 'randomized controlled trial' OR 'randomised controlled trial' OR 'randomized study' OR 'randomised study' OR 'randomised trial' OR 'randomized trial' OR 'randomly allocated' OR 'random assignment of intervention' OR 'randomization' OR 'randomisation' OR 'randomized experiments' OR 'randomised experiments' OR 'cluster randomised' OR 'cluster randomized' OR 'randomization procedure' OR 'randomisation procedure' OR random* OR 'random after matching' OR 'non-random' OR 'assigned randomly' OR 'assigned non-randomly' OR 'randomly assigned' OR 'non-randomly assigned' OR 'non-random study' OR 'permuted-block randomization' OR 'blocked randomization' OR 'stratified randomization' OR 'adaptive biased-coin randomization' OR 'urn randomization' OR 'covariate-adaptive randomization' OR 'response-adaptive randomization' OR 'outcome-adaptive randomization' OR 'permuted-block randomisation' OR 'blocked randomisation' OR 'stratified randomisation' OR 'adaptive biased-coin randomisation' OR 'urn randomisation' OR 'covariate-adaptive randomisation' OR 'response-adaptive randomisation' OR 'outcome-adaptive randomisation' OR 'simple randomization' OR 'restricted randomization' OR 'simple randomisation' OR 'restricted randomisation' OR 'randomized comparative trial' OR 'randomised comparative trial' OR 'randomized clinical trial' OR 'randomised clinical trial' OR 'assignment of treatments' OR 'assignment to group*' OR 'unit of assignment' OR 'group randomised' OR 'group randomized' OR individual OR 'individually randomised' OR 'individually randomized' OR 'matched at pre-test' OR matched OR 'two group*' OR 'pre-test' OR 'pre-test difference*' OR 'group equivalence' OR 'baseline equivalence' OR 'treatment group*' OR 'equivalent group*' OR 'matched-pair' OR 'matched pair' OR 'MP-RCT' OR 'experimental class*' OR 'trial subjects' OR 'intervention group*' OR 'treatment arm*' OR 'between-subjects design' OR 'experimental treatment' OR placebo OR 'no treatment' OR waitlist OR 'treatment vs treatment as usual' OR 'treatment as usual' OR 'placebo-controlled' OR 'post-hoc' OR 'superiority trial*' OR 'noninferiority trial*' OR 'single-blind' OR 'double-blind' OR 'triple-blind' OR unblinded OR 'equivalence trials' OR 'parallel-group*' OR crossover OR factorial OR explanatory OR pragmatic OR 'parallel trials' OR 'rotation design' OR 'multiple treatments' OR 'phase in' OR pipeline OR 'subgroup analys*' OR 'selection bias' OR 'allocation concealment' OR attrition) Limits applied

* Authors original, unpublished Table. Search limited by date (from 1 January 1980 to 31 December 2016) and language (English only).

For those studies located within the EEF and WWC, titles and abstracts were assessed by only one author as it was not possible to export them to an external repository for independent screening.

Studies which passed first level screening were then located and downloaded for full text screening. These PDFs were saved to an online library and stored under a unique study ID, the link was shared to authors and studies which met all the predetermined inclusion criteria were included in the final review. The first author (Paul Connolly) then assigned these studies to all authors and four additional graduate researchers. When studies could not be located via the usual methods (libraries, journals, inter-library loans) authors were contacted via email to request information.

The initial search for this systematic review was completed during 16–18 August 2015. Since then, three update searches have taken place. As there were no changes to the inclusion criteria or the objectives of the review, the search was re-executed in June 2016, April 2017 and again in August 2017, in accordance with the guidelines provided in the Cochrane Handbook (Higgins and Green 2011). All updates are included in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram

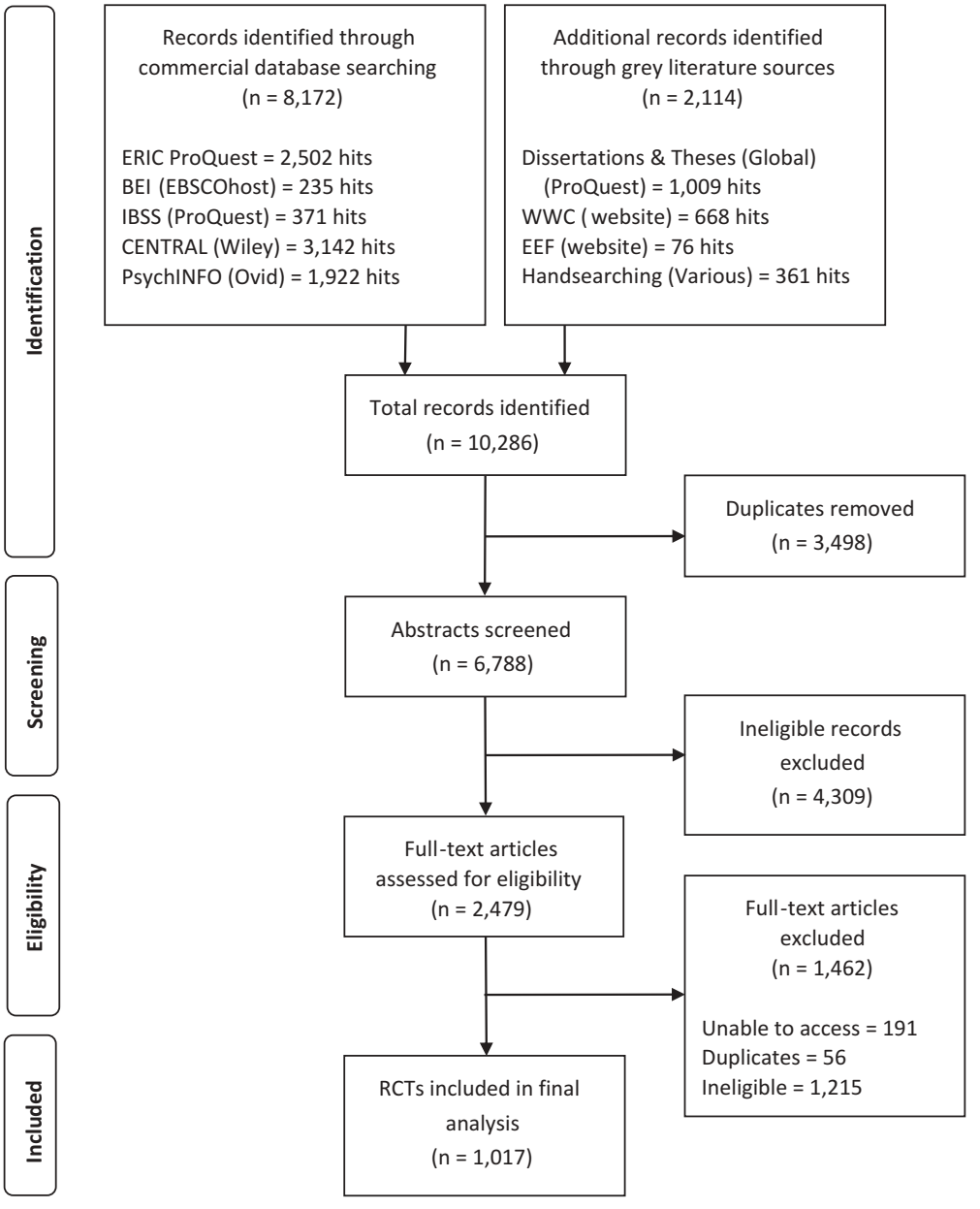

Figure 1. Flow chart summarising findings of search strategy*.
Source: Authors original, unpublished figure.

(see Figure 1) and new studies have been included in this current review (Liberati *et al.* 2009).

Data extraction and analysis

Data extraction sheets were designed by the first author (Paul Connolly) and piloted by another author (Ciara Keenan) using Microsoft Excel (2016). The following data were extracted from each included study:

- year of publication;
- location of study;
- type of educational institution that provided the focus for the intervention;
- whether the intervention was universally delivered or targeted at particular sub-groups of students;
- who delivered the intervention;
- the duration of the intervention;
- the primary and secondary outcomes focused on;
- whether the study used a simple or clustered RCT design;
- the total sample size;
- the number of clusters (where applicable);
- whether the studies reported statistically significant effects;
- whether the studies included a longitudinal element, beyond immediate post-test;
- whether the study included discussion regarding whether the findings could be generalized;
- whether the study included a qualitative process evaluation;
- whether the study included some reference to a theoretical framework in relation to describing the intervention being evaluated; and
- whether the study reflected upon existing theories in light of the findings from the trial.

The coding sheet, with instructions, is available from the first author on request.

The analysis consisted of a narrative synthesis, based upon descriptive statistics regarding the characteristics of the included studies. Within this, a particular emphasis was placed on summarising those characteristics of the included studies of relevance to the four key criticisms initially identified, as set out earlier.

Results

Figure 1 summarises the findings of the search strategy. The systematic search identified a total of 10,286 records that reduced to 6788 unique records once duplicates had been removed. These unique records were all screened and 2479 were identified for full-text assessment. This full-text assessment for eligibility resulted in a final sample of 1017 unique RCTs in education being identified for the period 1980–2016 inclusive.

The production of RCTs over time is illustrated in Figure 2. As can be seen, there has been a marked increase in the production of RCTs internationally over the last decade, with over three-quarters of all unique RCTs identified since 1980 (799 RCTs or 78.6% of the total) having been produced over the last 10 years (2007 to 2016). Overall, and as detailed in Table 3, just over half of all the RCTs identified (53.4%) were conducted in North America,

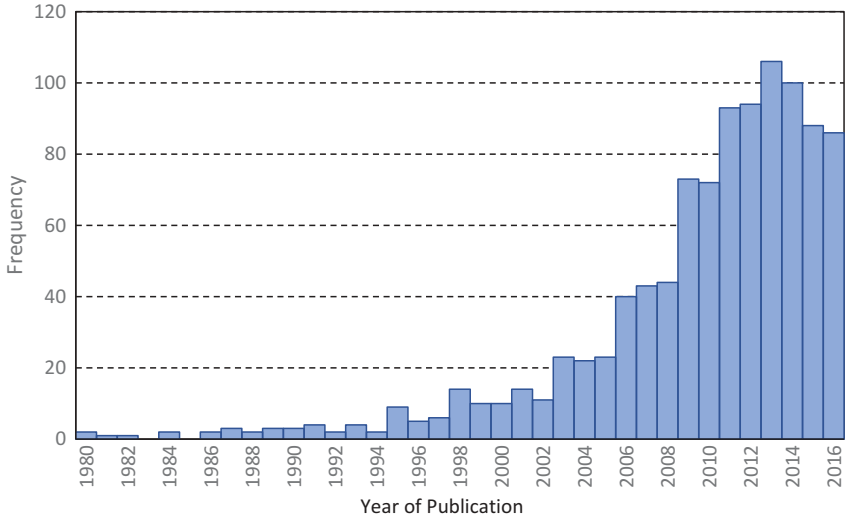

Figure 2. Number of randomised controlled trials in education completed internationally between 1980 and 2016*.

*Source: Authors original, unpublished figure.

Table 3. Characteristics of the randomised controlled trials*.

Characteristic	Type	Frequency	Valid %**
Type of trial	Single randomised	430	42.3
	Cluster randomised	586	57.7
	Not clearly stated	1	
	Total	1017	100.0
Total sample size***	0 to 50 participants	120	11.9
	51 to 250 participants	369	36.6
	251 to 1,000 participants	271	26.9
	More than 1,000 participants	248	24.6
	Not clearly stated	9	
	Total	1017	100.0
Location of trial	USA and Canada	543	53.4
	Europe (excluding UK and Ireland)	167	16.4
	UK and Ireland	131	12.9
	Asia	79	7.8
	Australia and New Zealand	55	5.4
	Africa	19	1.9
	Central and South America	18	1.8
	Multiple regions	5	0.5
	Total	1017	100.0

* Authors original, unpublished table.
** Columns may not sum to 100.0% due to rounding.
*** Mean = 1,251 (SD = 4,102); Min = 6; Maximum = 80,000.

with a little under a third (29.3%) conducted in Europe. Smaller numbers of trials can also be found elsewhere internationally, although it should be noted that these figures are likely to under-represent the actual total number of RCTs produced during this period due to the current search strategy being restricted only to articles/reports published in English. Interestingly, a little under half of the RCTs reported (42.3%) were simple RCTs, with 57.7% using clustered randomised designs (i.e. where the allocation of participants to control and intervention conditions is undertaken on a group basis, such as whole classes

Table 4. Characteristics of the programmes being evaluated*.

Characteristic	Type	Frequency	Valid %
Institution through which programme is delivered	Primary/elementary schools	340	33.4
	Middle/high schools	326	32.1
	College/university	198	19.5
	Preschool/kindergarten	79	7.8
	Multiple types of institution	61	6.0
	Special schools	13	1.3
	Total	1017	100.0
Main focus or primary outcome of programme	Physical health and well-being**	235	23.1
	Literacy/English	205	20.2
	Behaviour and social well-being	204	20.1
	Professional training	151	14.9
	Numeracy/Maths	70	6.9
	Range of academic outcomes	52	5.1
	Other school subjects	38	3.7
	Study-related skills	38	3.7
	Other	24	2.4
	Total	1017	100.0
Programme approach	Universal	680	66.9
	Targeted	337	33.1
	Total	1017	100.0
Programme delivered by	Regular teachers/lecturers	547	53.8
	External educators	279	27.5
	Mixture	120	11.8
	Other school/college employees	54	5.3
	Wider policy initiatives	16	1.6
	Not clearly stated	1	
	Total	1017	100.0
Duration of programme	Full academic year or longer	289	28.4
	Between half a term and a full term	253	24.9
	Up to half a term	251	24.7
	More than one term	148	14.6
	Single session	75	7.4
	Not clearly stated	1	
	Total	1017	100.0

* Authors original, unpublished table.** To be eligible, these programmes needed to include an educational component with the aim of improving at least one educational outcome.

or schools, rather than individually). Also noteworthy is the fact that around a quarter of all the RCTs identified were relatively large, with more than 1000 participants (24.6%).

The types of interventions evaluated through these 1017 RCTs are summarised in Table 4. As can be seen, about two thirds of all trials took place either in primary/elementary schools or middle/high schools (65.5%). A significant focus, for a little under half of the interventions (43.2%), was on educational programmes seeking to improve students' health and wellbeing (including those focusing on physical health and behaviour). Just over a third of the RCTs identified (35.9%) focused on interventions seeking to improve academic outcomes in some way; whether this be those focused specifically on literacy/English or numeracy/maths, or those focused on other school subjects and/or a range of academic outcomes. About two thirds of these interventions (66.9%) were universal in approach (i.e. whole-class or whole-school based) and with just over half of these (53.8%) delivered by the regular class teachers or lecturers. Just over a quarter of the interventions evaluated were facilitated by external educators (27.5%). It is also worth noting that just over half were delivered for more than half a term (53.3%) and a further 14.6% for more than one term. Moreover, it is notable that over a quarter of the interventions (28.4%) ran for one full year or longer.

Table 5. Characteristics of randomised controlled trial designs and approach to analysis*.

Characteristic	Type	Frequency	Valid %
Did the RCT include a qualitative process evaluation?	Yes, fairly well	208	20.6
	Yes, limited	102	10.1
	Yes, but not reported	71	7.0
	No	631	62.4
	Not clearly stated	5	
	Total	1017	100.0
Did the analysis of the RCT data include sub-group analyses?	Yes, at least some sub-group analysis	498	49.3
	No, just analysed sample as a whole	513	50.7
	Not clearly stated	6	
	Total	1017	100.0
Did the RCT find evidence of intervention effects?	Yes	816	80.8
	No	194	19.2
	Not clearly stated	7	
	Total	1017	100.0
Did the RCT include a longitudinal component Beyond immediate post-test?	Yes	461	45.5
	No	553	54.5
	Not clearly stated	3	
	Total	1,017	100.0
Did the analysis include some discussion of limitations to generalisation?	Yes	788	77.9
	No	223	22.1
	Not clearly stated	6	
	Total	1017	100.0
Did the study include some reference to theory?	Yes, discussed theoretical perspectives	353	34.8
	Yes, a descriptive theory of change	431	42.5
	No	231	22.8
	Not clearly stated	2	
	Total	1017	100.0
Did the study use the findings to reflect upon Implications for theory?	Yes	612	60.5
	No	399	39.5
	Not clearly stated	6	
	Total	1017	100.0

* Authors original, unpublished table.

Finally, further details on the characteristics of the RCT designs and their approaches to analysis are summarised in Table 5. As regards the methods used, studies were categorised in relation to whether they included a process evaluation or not. For the purposes of this review, a process evaluation was defined in terms of the use of qualitative methods to supplement, and provide insights into, findings from the quantified outcomes of the trial. These methods typically involved semi-structured interviews with participants and other stakeholders to document their experiences and perspectives but can also include naturalistic observations of the interventions as they are delivered. It can be seen that whilst nearly two thirds of the RCTs did not include a process evaluation component (62.4%), a little over a third did (37.7%), with one in five (20.6%) including a significant or notable process evaluation. Around a half of the studies (49.3%) included some sub-group analyses, where the effects of the intervention were not just analysed with regard to the sample as a whole but also in relation to whether the effects differed between subgroups of students (i.e. girls compared to boys or those from differing socio-economic backgrounds). As indicated, over three quarters of the RCTs produced evidence, in the form of statistically significant results, of effects of the intervention being evaluated (80.8%) and just over half studied the effects of the intervention beyond the immediate post-test period. In addition, just over three-quarters (77.9%) included some discussion of the limitations to generalisability of their findings. With regard to theory, just over three quarters of the RCTs identified

included some reference to theory when describing the intervention that was being evaluated (77.3%) and the majority included some reflection on existing theories in light of the findings of the RCT (60.5%).

Discussion

The findings outlined above help to bring some much-needed evidence to the increasingly trenchant debates surrounding the use of RCTs in education. In particular, they provide some challenge to the four key criticisms of RCTs outlined earlier. Thus, and firstly, they provide clear evidence to counter the claim that it is just not possible to do RCTs in education. As has been demonstrated, there now exist over 1000 RCTs that have been successfully completed and reported across a wide range of educational settings and focusing on an equally wide range of interventions and outcomes. Whilst there is a clear dominance of RCTs from the United States and Canada, there are significant numbers conducted across Europe and many other parts of the world. Many of these have been relatively large-scale trials, with nearly a quarter (248 RCTs in total) involving over one thousand participants. Moreover, a significant majority of the RCTs identified (80.8%) were able to generate evidence of the effects of the educational interventions under investigation. As noted earlier, these figures are likely to be under-estimates given the limitation of the present systematic review, with its restricted focus on articles and reports published in English. Nevertheless, the evidence is compelling that it is quite possible to undertake RCTs in educational settings. Indeed, across the 1017 RCTs identified through this systematic review, there are almost 1.3 million people that have participated in an RCT within an education setting between 1980 and 2016.

Secondly, there is some evidence to counter the criticism that RCTs ignore context and experience. Whilst they only constitute a minority of the trials identified (37.7%), there were 381 RCTs found that included a process evaluation component. Of course, this does mean that nearly two thirds of the RCTs found either did not include or failed to report a process evaluation element in their research designs. Moreover, and given the wide-ranging nature of this present study, it has not been possible to assess the quality or rigour of those RCTs that have included a process evaluation component. As such, it has not been possible to assess how well the process evaluation components have taken into account context and experience and/or engaged with, and successfully contributed to, the interpretation of the findings from the quantified outcomes. Nevertheless, there are sufficient numbers of RCTs that have been identified that have included a process evaluation component to suggest that it is possible for RCTs, as part of a mixed methods design, to include an emphasis on context and experience. Moreover, and in relation to the analysis of the quantified outcomes, it is encouraging to note that about half of the RCTs found (49.3%) included some consideration of the potentially differential effects of the intervention under study on differing subgroups of students, demonstrating some recognition of the potential for educational programmes to operate differentially across contexts and subgroups. However, it does also indicate that the other half of the RCTs have simply focused on the overall effects of the educational intervention in question and thus have not considered whether its impact varies across students.

Thirdly, there is more evidence to suggest that the RCTs produced within the time period have attempted to avoid the generation of universal laws of 'cause and effect'. Certainly,

those RCTs identified that have included at least some subgroup analyses would suggest a more nuanced approach amongst those conducting RCTs, that acknowledges that educational interventions are not likely to have the same effect across all contexts and all groups of students. Moreover, this is clearly evident amongst the majority of RCTs reported (77.9%) that included at least some discussion of and reflections on the limitations of the findings in terms of their generalizability. Such discussions were not querying the validity of RCTs per se but typically reflected a commendable level of critical reflexivity by the authors regarding how far their findings can be applied to other situations and contexts. However, and again, this does mean that nearly a quarter of RCTs to date have not recognised the need to qualify their findings in terms of stressing the difficulties of generalising to the wider population.

Finally, and in relation to the fourth criticism regarding the atheoretical nature of RCTs, this is also challenged to some extent by the findings presented above. A clear majority of RCTs that were reported included some discussion of the theory underpinning the interventions under investigation (77.3%). Moreover, a majority of RCTs (60.5%) also provided some reflections on the implications of their findings for theory. Whilst this is encouraging, the findings also suggest that a significant minority of RCTs – two out of every five (39.5%) – fail to use the opportunities provided by their study to engage in theory development.

Conclusions

Overall, the findings from this systematic review of RCTs undertaken in education 1980 – 2016 are mixed. On the one hand, there is clear evidence that it is possible to conduct RCTs in education, regardless of the nature of the education setting or of the particular type and focus of the intervention under consideration. This evidence is not just demonstrated by the 1,017 RCTs that this systematic review has identified, but also by the fact that over three-quarters of these RCTs have found evidence of intervention effects and nearly half have studied effects beyond immediate post-test. Moreover, the evidence also clearly refutes the claims that RCTs are, by their very nature, incapable of studying context or experience or that they always tend to generate simplistic laws of cause and effect and fail to make any meaningful contribution to theory. There are many examples of researchers conducting RCTs with a clear recognition of the need to incorporate a focus on context and experience through the inclusion of a process evaluation component to the research design and also the use of subgroup analyses. Moreover, many researchers reporting the findings of their RCTs are at pains to stress the difficulties of generalisation and also genuinely set out to use the RCT to test particular theories of human development and of change. Perhaps the key message from this present review is that it is quite possible to undertake an RCT, as part of a mixed method design, that is fully aware of and reflects the complexity of the social world.

However, and on the other hand, it is perhaps not surprising that criticisms of RCTs continue when nearly two thirds of RCTs in this period of time have not included a process evaluation component and where nearly half of them have not looked beyond the overall effects of the intervention in question for the sample as a whole. Similarly, it is difficult to challenge the view that RCTs promote a simplistic and atheoretical approach to educational research when nearly 40% of trials in this analysis have failed to reflect upon the implications of their findings for theory. This, however, is not an inherent weakness in the design of RCTs but rather should be considered as opportunities lost. With the increasingly widespread use

of RCTs in education, there is the growing expertise within the education research community to design and undertake more nuanced and sophisticated trials that explicitly seek to contribute to theory testing and development and that are acutely aware of the contingent and context-specific nature of educational interventions.

These represent clear challenges to researchers undertaking RCTs: to ensure that they include meaningful and rigorous process evaluations in their research designs; to ensure that their plans for analysing the quantified outcomes include a consideration of the potential impact of context in relation to exploring how intervention effects may vary for different subgroups of students and also in relation to different levels of delivery; and to engage much more centrally with underpinning theories. To support researchers seeking to rise to these challenges, there is a need to build upon the initial analysis and findings reported here. For example, it would be important to look more closely at those RCTs that have incorporated a process evaluation to assess how they have done this and what lessons can be learnt for integrating process evaluation methods with quantitative trial designs. Similarly, there is a need to explore what methods researchers have used to understand underpinning theories of change for particular educational interventions and to test these. In this respect, the increasing use of logic models in the design of educational interventions may also provide a helpful framework for specifying theories of change to then be tested. As explained elsewhere, in relation to RCTs in education, logic models provide a very useful framework for identifying and specifying: what *investments* are required with regard to any given intervention; what *activities* are needed to develop the intervention; what the specific *outputs* of these activities are, typically with regard to describing the key components of the intervention itself; and then how those outputs or components are believed to result in measurable improvements for the participants i.e. *outcomes* (see Connolly et al. 2017).

Finally, there is a need to increase our understanding of how researchers conducting RCTs have acknowledged and incorporated context within their analyses. With the growing number of RCTs now in existence, this is also where further systematic reviews and meta-analyses of substantive types of intervention holds out significant hope. Through the synthesis of data from a range of RCTs conducted across a variety of contexts, there is the genuine possibility of beginning to move on from the notion of 'what works' towards what works for whom, under what conditions and in what circumstances.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Paul Connolly http://orcid.org/0000-0001-9176-9592
Ciara Keenan http://orcid.org/0000-0003-0831-7347
Karolina Urbanska http://orcid.org/0000-0001-5063-4747

References

Bridges, D., P. Smeyers, and R. Smith, Eds. 2009. *Evidence-Based Education Policy: What Evidence? What Basis? Whose Policy?* Oxford: Wiley-Blackwell.

Campbell Collaboration 2016. *Campbell Collaboration Systematic Reviews: Policies and Guidelines* [Technical Report]. Retrieved from https://www.campbellcollaboration.org/library/campbell-collaboration-systematic-reviews-policies-and-guidelines.html

Centre for Reviews and Dissemination (CRD). 2009. *Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care*. Accessed 25.04.17. Centre for Reviews and Dissemination. York, UK: University of York.

Cohen, L., L. Manion, and K. Morrison. 2011. *Research Methods in Education*. 7th ed. London: Routledge.

Connolly, P., A. Biggart, S. Miller, L. O'Hare, and A. Thurston. 2017. *Using Randomised Controlled Trials in Education*. London: Sage.

Elliott, J. 2004. "Making Evidence-based Practice Educational". In *Evidence-Based Practice in Education*, edited by G. Thomas and R. Pring. Maidenhead: Open University Press.

Gorard, S., B. H. See, and N. Siddiqui. 2017. *The Trials of Evidence-Based Education: The Promises, Opportunities and Problems of Trials in Education*. London: Routledge.

Hammersley, M. 2007. *Educational Research and Evidence-Based Practice*. London: Sage.

Higgins, J. P. T., and S. Green, Eds.. 2011. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.1.0 [updated March 2011]: The Cochrane Collaboration, 2011 [accessed 10.04.17].Available from http://handbook.cochrane.org/.

Hodkinson, P., and J. Smith. 2004. "The Relationship between Research, Policy and Practice." In *Evidence-Based Practice in Education*, edited by G. Thomas and R. Pring. Maidenhead: Open University Press.

Liberati, A., D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. Ioannidis, and D. Moher. 2009. "The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies that Evaluate Health Care Interventions: Explanation and Elaboration." *PLoS Med* 6 (7): e1000100. doi:10.1371/journal.pmed.1000100.

Morrison, K. 2001. "Randomised Controlled Trials for Evidence-Based Education: Some Problems in Judging 'What Works'." *Evaluation & Research in Education* 15: 69–83. doi:10.1080/09500790108666984.

Richardson, W. S., M. C. Wilson, J. Nishikawa, and R. S. Hayward. 1995. "The Well-Built Clinical Question: A Key to Evidence-Based Decisions." *ACP Journal Club* 123 (3): A12–A12. doi:10.7326/ACPJC-1995-123-3-A12.

Rosenthal, R. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin* 86 (3): 638–641. doi:10.1037/0033-2909.86.3.638.

Sandieson, R. 2006. "Pathfinding in the Research Forest: The Pearl Harvesting Method for Effective Information Retrieval." *Education and Training in Developmental Disabilities*, 41 (4): 401–409.

Sandieson, R. 2017. Systematic Review Terms [Information on a Page]. Retrieved from https://pearlharvestingsearchthesaurus.wikispaces.com/Systematic±Review±Terms

Thomas, G., and R. Pring, Eds. 2004. *Evidence-Based Practice in Education*. Maidenhead: Open University Press.

Torgerson, C. J., and D. J. Torgerson. 2001. "The Need for Randomised Controlled Trials in Educational Research." *British Journal of Educational Studies* 49: 316–329. doi:10.1111/1467-8527.t01-1-00178.

Torgerson, D. J., and C. J. Torgerson. 2008. *Designing Randomised Trials in Health Education and the Social Sciences*. Basingstoke: Palgrave McMillan.