



A new methodological approach for evaluating the impact of educational intervention implementation on learning outcomes

Laura A. Outhwaite, Anthea Gulliford & Nicola J. Pitchford

To cite this article: Laura A. Outhwaite, Anthea Gulliford & Nicola J. Pitchford (2019): A new methodological approach for evaluating the impact of educational intervention implementation on learning outcomes, International Journal of Research & Method in Education, DOI: [10.1080/1743727X.2019.1657081](https://doi.org/10.1080/1743727X.2019.1657081)

To link to this article: <https://doi.org/10.1080/1743727X.2019.1657081>



© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 12 Sep 2019.



[Submit your article to this journal](#)



Article views: 334



[View related articles](#)



[View Crossmark data](#)

A new methodological approach for evaluating the impact of educational intervention implementation on learning outcomes

Laura A. Outhwaite^{a,b}, Anthea Gulliford^a and Nicola J. Pitchford^a

^aSchool of Psychology, University of Nottingham, Nottingham, UK; ^bCentre for Education Improvement Science, UCL Institute of Education, London, UK

ABSTRACT

Randomized control trials (RCTs) are commonly regarded as the 'gold standard' for evaluating educational interventions. While this experimental design is valuable in establishing causal relationships between the tested intervention and outcomes, reliance on statistical aggregation typically underplays the situated context in which interventions are implemented. Developing innovative, systematic methods for evaluating implementation and understanding its impact on outcomes is vital to moving educational evaluation research beyond questions of 'what works', towards better understanding the mechanisms underpinning an intervention's effects. The current study presents a pragmatic, two-phased approach that combines qualitative data with quantitative analyses to examine the causal relationships between intervention implementation and outcomes. This new methodological approach is illustrated in the context of a maths app intervention recently evaluated in a RCT across 11 schools. In phase I, four implementation themes were identified; 'teacher support', 'teacher supervision', 'implementation quality', and 'established routine'. In phase II, 'established routine' was found to predict 41% of the variance in children's learning outcomes with the apps. This has significant implications for future scaling. Overall, this new methodological approach offers an innovative method for combining process and impact evaluations when seeking to gain a more nuanced understanding of what works in education and why.



ARTICLE HISTORY

Received 14 December 2018
Accepted 18 July 2019

KEYWORDS

Implementation;
intervention; evaluation;
mixed-methods; education

The start of the twenty-first century saw the emergence, growth, and increased investment in evidence-based education (see Hanley, Chambers, and Haslam 2016; Thomas and Pring 2004; Hammersley 2007). Evidence-based education seeks to understand 'what works' and to date has shown a preference for experimental, quantitative, and post-positivist methodologies, particularly randomized control trials (RCTs; Connolly et al. 2017; Haynes et al. 2012; Torgerson and Torgerson 2003). However, this approach is substantially criticized (Connolly, Keenan, and Urbanska 2018). In particular, in establishing universal and replicable laws (Hodkinson and Smith 2004) RCTs are argued to be too reductionist for evaluation studies conducted in complex environments, such as schools (Biesta 2010). Specifically, the emphasis on statistical aggregation removes educational interventions and their outcomes from their situated context (Elliott 2001). To understand how a particular intervention works and under what circumstances, evaluation research designs need

CONTACT Nicola J. Pitchford  nicola.pitchford@nottingham.ac.uk  School of Psychology, University of Nottingham, University Park, Nottingham NG7 2RD, UK

This article has been republished with minor changes. These changes do not impact the academic content of the article.

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

to emphasize intervention implementation (Pawson and Tilley 1997; Slavin 2012) assessed through a process evaluation.

A recent systematic review of over 1000 RCTs conducted in education found only one in five studies evaluated intervention implementation and the methods used were frequently limited to descriptive qualitative data or critical reflexivity on the generalisability of the presented findings to other situations, contexts, and participant groups (Connolly, Keenan, and Urbanska 2018). In addressing these issues, educational evaluation research is growing in sophistication, by combining quantitative impact evaluations with qualitative methods, including implementation process evaluations (Wyse and Torgerson 2017; Humphrey et al. 2016). However, current methods do not easily connect results directly from the implementation process evaluation to the intervention outcomes, thereby limiting the conclusions that can be drawn.

As such, a more systematic, mixed-methods approach to evaluating implementation is needed (Oakley et al. 2006) that also affords statistical examination of the impact of the implementation process on learning outcomes (Peterson 2016; Shaffer 2011; Thomas 2016). The current study demonstrates a novel and informative methodology for examining intervention implementation within a determinant theoretical framework in the context of a recent RCT that evaluated a maths app intervention implemented across 11 primary schools (Outhwaite et al. 2018).

Maths app intervention

A growing evidence base demonstrates the educational benefits of app-based mathematics instruction for young children (Herodotou 2018; Pitchford 2015). In particular, a recent pupil-level RCT conducted in 11 schools found after a 12-week intervention period, children aged 4–5 years who used the maths apps either as an additional activity (treatment group) or instead of a small-group mathematical task (time-equivalent group) made significantly greater progress in mathematics compared to children who received standard mathematical practice (control group; Outhwaite et al. 2018). No main effect or interactions were found when School was entered as an independent variable in the quantitative analyses on mathematical progress, but understanding how the intervention was implemented across the 11 participating schools and how these school-level implementing factors may impact children's learning outcomes with the maths apps is vital. Specifically, we hypothesize that studying implementation factors will provide valuable insights for scaling the intervention, by taking into consideration the implementing teachers' experiences and expertise (Langley et al. 2009).

Defining implementation

Previous implementation evaluation research has focused on a single dimension of implementation; predominately fidelity i.e. the extent to which the intervention is delivered as intended, such as the structure and sequence of intervention activities or dosage (e.g. duration and frequency of the intervention; Berkel et al. 2011; Vignoles, Jerrim, and Cowan 2015). However, this approach has been criticized for over-simplifying implementation, which is a complex and multi-dimensional construct with several distinct but related dimensions (Durlak and DuPre 2008; Forman et al. 2009). Instead, there have been calls for an increased focus on other, multiple dimensions of implementation (Humphrey et al. 2016; Lendrum, Humphrey, and Greenberg 2016). For example, the TiDieR checklist, which encourages sufficient descriptions of interventions to support replication, and incorporates multiple definitions of implementation including fidelity, dosage, adaptations, and quality (Hoffmann et al. 2014). This study will focus on the quality of implementation and adaptations made by teachers when implementing a maths app intervention in their classroom that was examined through an RCT (Outhwaite et al. 2018).

Implementation quality refers to how well different aspects of the intervention are delivered (Durlak and DuPre 2008). For example, this study will focus on teachers' responsiveness to the delivery of the maths apps (O'Donnell 2008). In contrast, adaptations refer to the ways in which the intervention may be changed during implementation by teachers, to meet the needs of their specific

classroom circumstances and contexts (Hanley, Chambers, and Haslam 2016; Naylor et al. 2015). Adaptations are a natural process as implementing teachers are 'active modifiers' not 'passive acceptors' of a particular intervention (Rogers 2003). In this study, adaptations may include the logistical fit and timings of the maths app intervention delivery (Moore, Bumbarger, and Cooper 2013). Adaptations are highly likely when an intervention scales so it is critical to understand how adaptations in intervention implementation may impact on learning outcomes.

Implementation variability and intervention outcomes

In addition to developing a descriptive narrative of the intervention implementation in individual schools, it is also necessary to examine the relationship between implementation and outcomes (Elliott and Mihalic 2004). The evidence base linking implementation variability to educational intervention outcomes is sparse (Humphrey et al. 2016) and is largely situated in the health field. However, there are some promising results (e.g. Askill-Williams et al. 2013; Hansen et al. 2013). A recent study by Askill-Williams et al. (2013) generated an implementation index to quantitatively assess implementation quality, fidelity, and dosage of a mental health intervention implemented in primary schools. Results showed a significant relationship between high and medium rated intervention implementation and greater intervention outcomes over time, compared to low rated implementation. Similarly, Hansen et al. (2013) assessed the frequency and nature of teacher adaptations in the delivery of a school-based drug prevention programme using coded video observations. Teachers who made positive but fewer adaptations had a higher percentage of students that remained non-drug users compared to teachers who made more frequent adaptations, regardless of the positive, neutral, or negative rating given to the adaptations made. Together, these studies demonstrate how intervention implementation can be evaluated and examined in relation to intervention outcomes. They highlight the need to develop and refine innovative methodological approaches for evaluating intervention implementation to provide high-quality research standards.

Determinant theoretical framework

Determinant frameworks aim to identify the barriers and enablers for successful implementation that influence intervention outcomes (Nilsen 2015). Determinant frameworks are typically multi-level; the intervention is placed at the centre surrounded by different influencing factors at different system levels (e.g. Domitrovich et al. 2010). In the context of app-based mathematical interventions (e.g. Outhwaite et al. 2018; see Figure 1), influencing factors on the intervention outcomes at the individual-level may include the child's working memory capacity (Cragg et al. 2017; Gathercole and Alloway 2006), socio-economic status (SES; Anders et al. 2012; Kalaycioglu 2015) and English as an additional language (EAL; Strand, Malmberg, and Hall 2015). Previous research evaluating the maths app intervention at the focus of this study found children's learning outcomes with the apps were not influenced by their SES or EAL status and children with weaker memory skills made greater progress compared to children with stronger memory skills (Outhwaite, Gulliford, and Pitchford 2017). At the school-level, factors may include how the intervention is implemented by teachers in their classroom. Specifically, it is suggested technology alone will not lead to learning, but is dependent on how the technology is integrated into the classroom environment (Beach and O'Brien 2015; Couse and Chen 2010). Hence, differences in implementation across participating schools may influence learning outcomes associated with the maths app intervention (Cook and Odom 2013; Humphrey et al. 2016). At the macro-level, factors may encompass senior teaching leaders' beliefs and values regarding the use of tablet devices in their school (Blackwell, Lauricella, and Wartella 2014). The current study focused on school-level factors and utilized mixed qualitative and quantitative methods to explore how implementation of the maths app intervention is associated with children's learning outcomes.

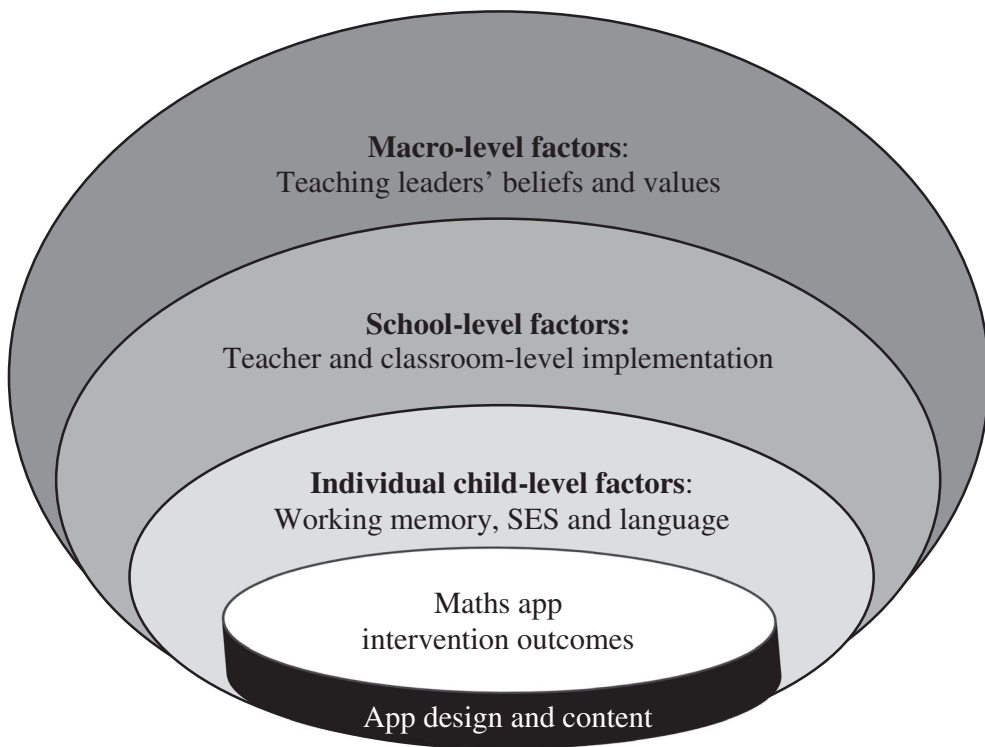


Figure 1. Multi-level, determinant framework outlining factors that may influence maths app intervention outcomes.

Research questions

This mixed-methods study addressed two questions in relation to Outhwaite et al.'s (2018) RCT findings: (1) how was the maths app intervention implemented in participating schools? (2) Is there a relationship between identified implementation themes and children's learning outcomes with the maths apps? To address these two distinct but related questions two phases of analysis were conducted.

Phase I adopted an inductive, bottom-up, thematic analysis approach to identify detailed themes of how the maths apps were implemented by individual schools. To achieve the insights needed, narrative direct observations of the maths app intervention sessions and semi-structured, self-report interviews with participating teachers were conducted. In phase II, a structured judgements approach was employed (Clarke 2004) to examine the relationship between implementation themes identified in phase I and children's learning outcomes with the maths app intervention (Outhwaite et al. 2018).

Phase I: understanding variation in implementation

Methods

RCT design

The implementation evaluation reported here builds on the quantitative data analysis of learning outcomes in Outhwaite et al. (2018). This pupil-level RCT evaluated the effectiveness of a new maths app intervention compared with standard mathematical practice with children aged 4–5 years. Within each class of 11 participating schools, children were randomly allocated to one of three groups, see Table 1. A treatment group used the maths app intervention in addition to regular mathematics instruction, a time-equivalent treatment group received the maths app intervention instead of a daily small group-based mathematics activity, so time spent learning mathematics was equivalent to children in a control group who continued to receive standard teacher-led mathematical practice.

Table 1. Summary of the research design and results reported in Outhwaite et al. (2018).

Mathematics activities and results	Treatment group	Time-equivalent treatment group	Control group
Maths app intervention	✓	✓	
Small group-based maths instruction	✓		✓
Whole class embedded maths activities	✓	✓	✓
Total time learning maths	Additional	Typical	Typical
Within-group effect size of maths progress (Cohen's <i>d</i> , 95% CI)	.78 (.42–1.14)	.65 (.29–1.00)	.47 (.13–.82)

The intervention was implemented by teachers for 30 minutes per day across 12 consecutive weeks during the 2016 Summer Term (April–June). Prior to the intervention, teachers were trained in how to implement the apps by the research team including an experienced Early Years teacher. A teacher manual provided additional implementation support and further details of the study protocols. This was intended to maximize consistency across participating schools. Schools had autonomy over the timing and logistics of the implementation to meet the needs of their individual classroom and school routines.

All children were assessed on a standardized measure of mathematical ability, before and immediately after a 12-week intervention period. Results showed children who used the maths apps made significantly greater progress in mathematics compared with standard mathematical practice only. There was no significant difference in learning outcomes between the two forms of app implementation. Further details on the methods and procedures, including intervention content can be found in Outhwaite et al. (2018). The current study expands on this work by taking an in-depth focus on implementation, as this will add significant insights for future scaling of this intervention.

The maths apps used in this study were developed by an educational not-for-profit organization, to which the independent research team had no vested interest in demonstrating the intervention was effective. The School of Psychology Ethics Committee at the University of Nottingham granted ethical approval for the study. Opt-in parental consent was obtained for all participating children in line with the British Psychological ethical guidelines.

Methodological position

The current study adopted a pragmatic, relativist, mixed-methods approach (Tashakkori and Teddlie 2010). Phase I was primarily qualitative (Creswell et al. 2006) that employed an inductive, bottom-up approach to identify implementation themes in the data (Punch 2013; Nilsen 2015). This was chosen due to the sparse literature on the implementation of app-based mathematics instruction (Morse 1991). Two sources of qualitative data were utilized to support implementation evaluation: (1) narrative direct observations made by the researcher (first author) and (2) self-report, semi-structured interviews, with implementing teachers. Data collected from these sources were assimilated to maximize internal validity (Humphrey et al. 2016) and provide a detailed holistic account of the intervention implementation experience within the practical constraints of the study (Clarke 2004).

Narrative direct observations

To illuminate how the maths app intervention implementation varied across the 11 participating schools, exploratory, direct observations were conducted in-situ by the first researcher (Robson and McCartan 2011; Hansen 2014; Humphrey et al. 2016). Observations were conducted during a school visit made approximately halfway through the 12-week intervention period. Both intervention treatment groups (see Table 1) were observed once. The researcher conducted the observations as a marginal participant (Robson and McCartan 2011), whereby they were completely accepted in the classroom environment but remained passive while observing the maths app intervention session. The researcher was positioned close to the children using the maths apps, so that it was possible to observe children's interactions with the technology without disrupting the session.

The observations were semi-structured and guided by interests in implementation quality and adaptations (Creswell and Creswell 2018). As the observations were exploratory, a flexible design was followed (Irwin and Bushnell 1980). Initial areas of observation interest included how children used the iPads, how the intervention organized in individual classrooms, and how teachers assisted children using the technology. Additional observations that were not initially anticipated were also recorded. Direct observation fieldwork notes were recorded during the observations by the researcher. This pragmatic approach was taken to develop a full and descriptive narrative of the maths app intervention implementation (Robson and McCartan 2011). This level of detail and complexity can frequently be lost in more structured approaches to observational methods (Creswell and Creswell 2018).

Self-report interviews

To further understand how the maths apps were implemented in the 11 individual school contexts, free flowing, semi-structured interviews were conducted with participating teachers during school visits made by the first author. The interviews were guided by the following questions: (1) How have you found implementing the maths apps? (2) What have you found challenging about implementing the maths apps? (3) What successes have you had with the maths apps?

This protocol was a guiding instrument only (Cohen, Manion, and Morrison 2007). The self-report interviews were grounded as an authentic experience, whereby the researcher followed the lead of the teacher. This helped to maintain the natural, free flowing nature of the interview conversation. Fieldwork notes were recorded by the researcher during the interview. Interviews were not audio recorded, as the data collected did not require extensive interpretation and re-construction, as is sometimes seen in qualitative research methods (Creswell and Creswell 2018). Instead, the interview data collected focused on explicit and surface level meaning of teachers' practical responses (Cohen, Manion, and Morrison 2007).

Qualitative validity and reliability

To ensure the validity (authenticity) and reliability (trustworthiness) of the research findings and consequent interpretations (Brantlinger et al. 2005; Creswell and Creswell 2018) the following five measures were taken.

Member checking. When conducting the self-report interviews participants' responses as recorded by the researcher were checked back with the participant throughout the duration of the interview. This ensured accuracy and verification of the information.

Thick, detailed, description. Where possible sufficient descriptions were recorded in the fieldwork notes to support the researcher's surface level interpretations and conclusions.

Familiarity with the participating organization. A staff member from each participating school met with the research team at a recruitment event (Outhwaite et al. 2018). The researcher also visited all participating teaching staff prior to study commencement and was in regular communication with participating schools throughout the duration of the study. This also helped to establish a rapport between the researcher and all of the participating teachers.

Addressing demand effects. Prior to the self-report interviews participants were encouraged to be honest and were assured that there were no right or wrong answers and that we, as researchers independent from the app development wanted to understand their opinions and experiences of implementing the maths app intervention. This step is vital as self-report interviews are frequently vulnerable to bias as participants do not wish to be negatively perceived by the researcher (Humphrey 2013).

Audit trail. Detailed records of all data collected, coded, displayed, synthesized, and interpreted were kept securely throughout the study.

Participants

Table 2 summarizes the descriptive data for participating teachers. Descriptive data were collected through a self-report, end-of-project feedback questionnaire, and was available for eight of the 11 participating schools.

School context

Originally 12 schools took part in the study from across Nottingham and Nottinghamshire in the East Midlands, United Kingdom (see Outhwaite et al. 2018), a geographical area with high levels of educational underachievement relative to other areas of the UK (Ofsted 2013). One school was not available at post-test due to a field-trip so was excluded from the final sample. It was not possible to follow-up on this school as post-testing took place during the last week of the school year.

The final sample of 11 participating schools represented a range of socio-economic and multicultural backgrounds and had a range of characteristics and Ofsted report ratings (Ofsted n.d.). Ofsted inspection reports assess schools as outstanding, good, requires improvement, or inadequate and can be used as a proxy measure of school quality (Gambaro, Stewart, and Waldfogel 2015; Schagen et al. 2005). These judgement ratings are based on direct, structured observations of the school and teaching environment conducted by a team of Ofsted inspectors typically every six years. Although inspectors have clear and detailed guidance on inspection criteria, ratings are based on the discretion of inspectors, which may vary (Schagen et al. 2005). The proportion of children eligible for pupil premium or free school meals (additional government funding for children from families earning below a certain threshold and those in local authority care) as reported in the school's most recent Ofsted report was used to indicate, tentatively, the SES of the school population. For example, a school described as having an SES 'below national average' was reported to have above national average proportions of children eligible for these funds in the relevant Ofsted report. Table 3 summarizes the profile of each school in the final sample.

Results

Qualitative data analysis

Table 4 summarizes data handling and analytical procedures employed in this study (Creswell and Creswell 2018; Clarke 2004). In phase I, qualitative data analysis was conducted to generate a narrative account of the intervention implementation from direct observations and self-report interviews to illuminate variations in the intervention implementation across the 11 participating schools. This was achieved by identifying implementation themes from the observation and interview datasets using an inductive, bottom-up, approach (Creswell and Creswell 2018).

Open-coding was used with the observation and interview datasets to identify units of analysis (Cohen, Manion, and Morrison 2007). New codes were generated until coding was complete; the

Table 2. Descriptive data for participating teachers.

Characteristic	Descriptive data
Number of teachers per early years class (min–max)	1–3
Number of years teaching (mean [SD], min–max)	9.92 (7.86), 2–25
<i>Teaching role (total frequency)</i>	
Senior leaders (e.g. head teacher with teaching duties)	1
Teachers (including early years leaders)	9
Teaching assistants	3
<i>Technology experience (total frequency)</i>	
Self-rated 'Experienced'	10
Self-rated 'Little experience'	3

Table 3. Profiles of the final sample of 11 participating schools.

School	Ofsted rating	Indicated SES profile	Other school characteristics
A	Good (2014)	Above national average	School A was smaller than average sized, mixed gender primary school (aged 4–11 years). The majority of children were White British and the number of children with SEN was below the national average.
B & C	Outstanding (2013)	Above national average	Schools B & C are two campus sites of the same large mixed gender primary school. Children attend one site only. The majority of children were White British, with a small minority of children from other ethnic groups. The number of children with SEN was below the national average.
D	Requires improvement (2013)	Above national average	School D was an average sized, mixed gender primary school. The majority of pupils were White British and children with SEN were below the national average.
E	Good (2013)	Below national average	School E was a smaller than average sized, mixed gender infant school (ages 4–7 years). The majority of children were White British and the number of children with SEN was below the national average.
F	Good (2012)	Below national average	School F was a larger than average sized, mixed gender Academy primary school. Children were mostly White British and the number of children with SEN was above the national average.
G	Good (2013)	Above national average	School G was an average sized, mixed gender primary school and the majority of pupils were White British. The number of children with SEN was below the national average. This school was not available at post-test due to a fieldtrip so were not included in the final sample.
H	Outstanding (2010)	Above national average	School H was a larger than average, mixed gender infant school. Children came from a wide range of ethnic backgrounds; the largest proportion was Indian or Pakistani. The number of children with EAL and SEN was above the national average.
I	Good (2011)	Above national average	School I was a larger than average sized, mixed gender Academy primary school. The majority of children were from White British, Asian or Asian British backgrounds. The number of children with EAL was above the national average and children with SEN were in line with the national average.
J	Good (2014)	Above national average	School J was a larger than average, mixed gender primary school. The number of children with SEN and children from ethnic backgrounds with EAL was above average.
K	Good (2010)	Not stated	School K was a larger than average, mixed gender primary school. The majority of children were of White British, Indian or Pakistani background. Approximately 25% of children spoke EAL and the number of children with SEN was below average.
L	Outstanding (2011)	Below national average	School L was an average sized, mixed gender Academy primary school and the majority of children were from ethnic minority background. The number of children with EAL and SEN was above the national average.

Note: SEN: special educational needs; EAL: English as an additional language; SES: socio-economic status.

dataset was saturated and all data was accounted for (Saldaña 2015). To ensure the codes were exhaustive, exclusive and consistent data codes were refined through three rounds of coding (Creswell and Creswell 2018; Saldaña 2015). The coding process generated semantic themes by focusing on explicit and surface level meaning of the data (Clarke and Braun 2014; Cohen, Manion, and Morrison 2007). Following three rounds of coding, four implementation themes were identified; (1) 'teacher support', (2) 'teacher supervision', (3) 'intended implementation', and (4) 'established routine'. Figure 1 summaries the thematic map for the four identified intervention implementation themes.

'Teacher support'

'Teacher support' encompassed four ways teachers assisted children to use the maths app intervention. These included (1) providing technical support to help children use the iPad device, such as ensuring headphones were correctly plugged into the device, (2) providing behavioural management, for example re-focusing children on the maths app activity when they became distracted or restless, (3) giving encouragement to support children to persevere with a maths app activity and (4) providing guidance when children needed assistance in understanding the app instructions.

Table 4 . Qualitative data analysis process (Creswell and Creswell 2018; Clarke 2004).

Stage	Description
<i>Phase I: Understanding implementation</i>	
Data Preparation	Original fieldwork notes photocopied and sorted securely. Original fieldwork notes anonymized (i.e. school name removed). Original fieldwork notes typed up and organized ready for analysis.
Data Familiarization	The researcher read through the raw observation and interview datasets twice to gain a general and overarching sense of the data.
Data Coding	Inductive, bottom-up approach with open coding. A series of informal initial codes were established by reviewing the datasets. Codes refined as researcher progressed through the data sets. Two cycles of data coding were conducted to ensure the data was saturated and data codes were exclusive, exhaustive, and consistent. A further cycle of coding was conducted to winnow data codes and ensure codes were exclusive.
Data Display	Thematic map generated incorporating the four implementation themes (see Figure 2).
<i>Phase II: Relationship between implementation & intervention outcomes</i>	
Interpretative Synthesis	Structured Judgement Methodology: Aggregated qualitative data for the four implementation themes presented as descriptive summaries for each school to convey individual experiences of implementing the maths app intervention.
Systematic Quantitative Framework	Structured Judgement Methodology: systematic quantitative framework designed and implemented (3-point Likert scale) to assess aggregated qualitative data. This step aimed to move data from individual description to general explanations (Clarke 2004).

'Teacher supervision'

'Teacher supervision' described the extent to which teaching staff actively supervised children using the maths app intervention and was characterized in two ways; (1) constant and (2) consistent throughout the intervention session.

'Intended implementation'

'Intended implementation' incorporated four ways in which the maths app intervention was implemented as intended, as outlined in the teacher manual and the training session that teachers received prior to the intervention commencement (Outhwaite et al. 2018). These features included (1) children using their own iPad device, (2) with headphones and (3) accessing the maths app content within their own in-app profile and (4) within a calm classroom environment.

'Established routine'

'Established routine' comprised five actions which supported the maths app intervention to be successfully embedded into the daily classroom schedule. These actions included (1) implementing the intervention at a consistent time each day, (2) having a dedicated member of staff whose responsibility it was to implement the intervention, (3) having well organized equipment, for example, colour coding the iPad devices so that they were easily identifiable by children, (4) having a dedicated space within the classroom and (5) a seating plan where children used the maths apps.

Discussion

Phase I generated a detailed, qualitative, understanding of the implementation of the maths app intervention by the 11 schools participating in the RCT reported by Outhwaite et al. (2018). Based on narrative direct observations and self-report interviews, a thematic analysis identified four intervention implementation themes; (1) 'teacher support', (2) 'teacher supervision', (3) 'intended implementation' and (4) 'established routine'. Collectively, these themes describe the implementing teachers' experiences (Peterson 2016) and classroom context in which the maths app intervention was situated (Biesta 2010; Humphrey et al. 2016). These methods are similar to those used in previous implementation process evaluations (Connolly, Keenan, and Urbanska 2018).

However, these qualitative themes do not capture the extent to which implementation varied across the 11 participating schools, nor do they elucidate how the extent of variation across these themes might be associated with children's learning outcomes with the maths apps. A key aspect of implementation science involves examining the variability in intervention implementation across contexts and learning from this variability (Peterson 2016). Phase II of this study sought to achieve this through examining the relationship between the four intervention implementation themes and children's learning outcomes with the maths apps.

Phase II: relationship between implementation & intervention outcomes

Methods

To examine the relationship between the four intervention implementation themes and children's learning outcomes with the maths apps (Outhwaite et al. 2018) a structured judgement approach was adopted (Clarke 2004). Structured judgement methods aim to move from single, particular descriptive data, such as the implementation themes identified in phase I, to general explanations based on aggregated data. Disciplined and structured judgements on aggregated qualitative data can be achieved by applying a systematic, quantitative, framework (Clarke 2004).

Aggregated qualitative data

To explore the relationship between the implementation of the maths app and learning outcomes it was necessary to gain further insight into the implementation conditions in each of the 11 participating schools. The individual schools' experiences were therefore reviewed against each of the four implementation themes.

To achieve this, the narrative direct observation and self-report interview data sources were assimilated into a descriptive body of data per school (see Table 4, interpretative synthesis) that were structured around the four implementation themes. These descriptive summaries based on the aggregated observation and interview data conveyed the individual schools' experiences of implementing the maths app intervention. An example descriptive summary is illustrated in Table 5.

Systematic quantitative framework design

To make disciplined and structured judgements on the aggregated observation and interview data, a 3-point Likert rating designed to assess the variability within the implementation themes across each school was utilized. Scorings of high (score 3), medium (score 2), and low (score 1) were applied. Necessary features for high (scored 3), medium (scored 2), and low (scored 1) ratings were identified based on the code definitions (see thematic map; Figure 2) and are outlined in detail in Table 6. As such, the 3-point Likert rating scale was relative (rather than absolute) across the final sample of 11 participating schools.

Systematic quantitative framework implementation procedure

When assigning the 3-point Likert scale ratings the researcher (first author) read through all of the descriptive summaries for each implementation theme and allocated an initial rating. This procedure was repeated to ensure consistent and fair ratings across the whole sample. As the researcher collected and coded the original dataset, an additional independent researcher, who was not involved with the original RCT or participating schools, applied the same rating procedure. A high inter-rater reliability was established with an average 90.6% agreement across the four themes (Miles and Huberman 1994). Any discrepancies in ratings were addressed through discussion between the two researchers until an agreement was reached.

Table 5. An example descriptive summary synthesizing observational and interview data for one participating school.

Theme	Synthesized observational and interview data
Teacher Support	In the observed sessions, the student teacher supported the intervention implementation through providing behaviour management by reminding children to be quiet and to re-focus them on the task. They also gave clear finishing instructions at the end of the sessions and provided technical support for closing the apps and putting away the iPads. In the staff interviews, the teaching assistant emphasized that the teachers provide behaviour management support (e.g. staying on task). Teachers also help with technical support (e.g. making sure the headphones are working correctly) and with encouragement (e.g. supporting children to repeat quiz activities where necessary). In the interview, the teaching assistant commented that the need for encouragement and guidance reduced as children became more independent and accustomed to the apps.
Teacher Supervision	In the observed sessions, the student teacher sat with the small group of children using the maths apps and had a constant and consistent presence with them. The teaching assistant also kept an overarching eye over the session.
Intended Implementation	In the observed session, all children were working within their own profile on their own iPad with headphones. Children were not distracted by the other activities in the wider classroom unit. The overall atmosphere was calm.
Established Routine	During the interview, the teaching assistant commented that daily implementation of the maths apps was the responsibility of the student teacher. They said the biggest challenge was establishing the routine at the start of the intervention period, particularly as the iPads are shared with the whole school. However, they commented that the intervention has slotted well into the school day; the children used the maths apps at a dedicated time each day. During the observed session, children were sat around a table in the main classroom unit in a dedicated seating plan. At the end of the observed session, the student teacher put away the equipment. However, in the teacher interview, the teaching assistant commented that one child has started taking on this responsibility.

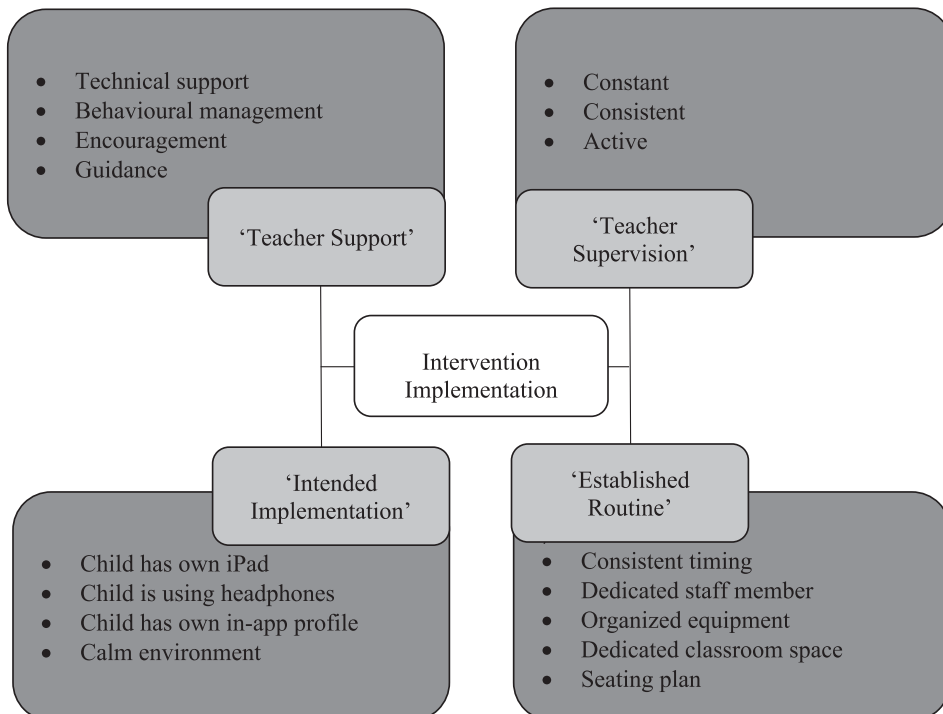
**Figure 2.** Phase I: thematic map summarizing the four implementation themes identified.

Table 6. Definitions for high, medium, and low ratings with guiding questions for each of the implementation themes.

Implementation theme	Guiding question for ratings	High, medium, and low features
'Teacher support'	To what extent did teachers give <i>support</i> to children using the maths apps?	High (scored 3): 3–4 types of support were reported (e.g. behavioural management, technical support, and encouragement and guidance) and were consistent throughout the intervention session. Medium (scored 2): 2 types of support were reported, which were intermittent throughout the intervention session. Low (scored 1): 1 (or less), type of support was reported and was intermittent throughout the intervention session.
'Teacher supervision'	To what extent did teachers <i>supervise</i> children using the maths apps?	High (scored 3): Constant supervision from consistent, dedicated member(s) of staff was reported throughout the intervention session. Medium (scored 2): Teaching staff were consistently present but not actively supervising children. Supervision from additional staff may have been reported, but this was not constant or consistent throughout the intervention session. Low (scored 1): Little (e.g. overarching eye from one teacher) to no supervision was reported.
'Intended implementation'	To what extent was the <i>intervention implemented as intended</i> ?	High (scored 3): The maths app intervention was implemented as intended, as instructed by the teacher manual and teacher training session (e.g. children used their own iPad, children accessed app content in their own profile, correct equipment including headphones were used, and there was a calm classroom environment) and was consistent throughout the intervention session. All aspects of the definition must be present. Medium (scored 2): 2–3 of the implementation criteria were correctly followed, but were intermittent throughout the intervention session. Low (scored 1): Little to no (1 or less), implementation criteria were correctly followed.
'Established routine'	To what extent was a <i>daily routine</i> established?	High (scored 3): 4–5 actions for establishing a consistent, daily routine were reported (e.g. member of staff assigned responsibility of the intervention, dedicated space for children to use the apps with a seating plan, intervention timing consistent throughout the week, and well-organized logistics). Medium (scored 2): 2–3 actions for establishing a daily routine were reported, but consistency may be varied. Low (scored 1): Little to no (1 or less) evidence that a consistent, daily routine was established.

Results

Preliminary Ofsted and SES analyses

A series of spearman's rho correlations showed no significant relationships ($p > .05$) between the schools' Ofsted ratings (see Table 3) and ratings for each of the four intervention implementation themes; 'teacher support' ($r^s = -.31$), 'teacher supervision' ($r^s = -.38$), 'intended implementation' ($r^s = -.39$), and 'established routine' ($r^s = .36$). Furthermore, no significant relationships ($p > .05$) were observed between school SES (see Table 3) and each of the four intervention implementation themes; 'teacher support' ($r^s = .33$), 'teacher supervision' ($r^s = .20$), 'intended implementation' ($r^s = .00$), and 'established routine' ($r^s = -.25$).

Learning gains

Within-group effect sizes (Cohen's d) were used to measure the extent of learning gains following the 12-week intervention period and were calculated for the treatment group and time-equivalent group, collapsed across the final sample of 11 participating schools. Although there was a difference in overall exposure to learning mathematics (see Table 1), an independent samples t-test found no significant difference in progress in mathematics, as indicated by the within-group effect sizes between the two maths app treatment groups, $t(20) = .84$, $p = .808$. This is probably because time spent with the maths app intervention was equivalent across the two treatment

groups. As such, it was deemed suitable for the effect sizes to be collapsed across the two maths app treatment groups, which produced an overall learning gains effect size per school. This school level implementation effect size captured the magnitude of progress from pre- to post-test per school.

Implementation themes

To examine the relationship between the implementation themes and children's learning outcomes in response to the maths app intervention, spearman's rho correlations were conducted (Outhwaite et al. 2018). Results showed a strong, positive, significant correlation between the implementation theme 'established routine' ($M = 2.27$, $SD = .79$) and learning outcomes ($M = .77$, $SD = .36$, $r^2 = .73$, $p = .011$). No other significant correlations with learning gains ($p > .05$) were identified: 'teacher support' ($M = 2.18$, $SD = .87$, $r^2 = .23$), 'teacher supervision' ($M = 2.00$, $SD = .89$, $r^2 = -.05$), and 'intended implementation' ($M = 2.55$, $SD = .69$, $r^2 = .02$).

Despite the small sample size ($n = 11$) an exploratory linear regression was conducted to explore the extent to which 'established routine' predicted learning outcomes. Results showed 'established routine' significantly predicted children's learning outcomes with the maths apps ($\beta = .29$, $p = .035$), accounting for 41% of the observed variance, $R^2 = .41$, $F(1,9) = 6.13$, $p = .035$.

Discussion

Phase II examined the relationship between intervention implementation and children's learning outcomes with the maths apps. Following a structured judgements approach, qualitative data from the narrative direct observations and self-report interviews was aggregated for of the four implementation themes identified in phase I. This conveyed the individual maths app intervention implementation experiences in the final sample of 11 participating schools. A 3-point Likert scale systematic quantitative framework was utilized to make structured judgements about the aggregated qualitative data for each school. These ratings were then correlated with children's learning gains with the maths apps.

Results showed 'established routine' was the implementation theme most closely related to learning outcomes with the maths app intervention, as a strong, positive and significant correlation was found. When entered into an exploratory linear regression model, 'established routine' ratings accounted for 41% of the observed variance in learning outcomes. With a small sample size ($n = 11$), this tentatively indicates that schools that had a well-established daily implementation routine made the most progress in mathematics over time with the intervention. Thus, even with an app-based intervention that required minimal input from teachers in terms of delivery, classroom practice is crucial in determining the success of the intervention. This is essential to consider in the further scaling of this intervention and supports the assertion that how technology is integrated into the school environment is critical to its success (Beach and O'Brien 2015; Couse and Chen 2010). Furthermore, this evidence echoes direct instructional theory, which emphasizes the structural conditions for learning with repeated rehearsal and reduced distraction in the environment (Kirschner, Sweller, and Clark 2006).

No relationship was found between rating for the four implementation themes and Ofsted ratings, a proxy measure for school quality (Gambaro, Stewart, and Waldfogel 2015; Schagen et al. 2005), or schools' SES. As such, the significant association between 'established routine' and learning gains with the maths app intervention reflects the influence of other factors. For example, teacher's perceptions of educational technology are associated with positive uptake and integration of technology in the classroom (Blackwell, Lauricella, and Wartella 2014). To assess this potential source of influence, further research is needed to expand the determinant framework reported here (Figure 1) to include measures at the macro-level where teacher's perceptions and beliefs of educational technology would be situated.

Other implementation themes

No statistically significant relationships were observed between children's learning gains and the other three implementation themes. For 'teacher support' and 'teacher supervision', these results may be expected due to the child-centred nature of the maths app software. In particular, the maths apps are age-appropriate (Kucirkova 2014) and are grounded in instructional psychology (Kirschner, Sweller, and Clark 2006; Gray 2015) and learning science theory (Hirsh-Pasek et al. 2015). This high-quality app design may enable children to access learning activities without a specific need for direct 'teacher support' or supervision. This supports the assertion that app technology can provide effective maths instruction without additional, time-consuming, teaching demands (Hilton 2016; Kucian et al. 2011).

Future directions

Future research needs to consider two issues to enhance understanding of how variation in the implementation of education interventions may impact on learning outcomes. Firstly, the momentary nature of the narrative direct observation and self-report interview methods used in this study afford a time limited understanding of user's experience of the maths apps. Some fluctuations in daily practice are anticipated, which poses a threat to the internal validity of the current findings. Although the assimilation of two data sources helped to address this issue, it is important to consider this caveat when interpreting the current findings. Furthermore, gaining an understanding of user experience is distinct from treatment integrity measures that formally assess implementation fidelity (Humphrey et al. 2016). As such, future implementation evaluation studies should consider expanding the definition of implementation and examine how these different factors may influence the integration and success of educational interventions. Secondly, the determinant framework outlined above should include child engagement with the intervention at the individual level. This measurement of individual child engagement could help illuminate its association with the maths app-related learning gains and understand more about the universal reach and application of the maths app intervention to maximize the development of children's early mathematical skills.

Furthermore, this determinant framework (see [Figure 1](#)) is suited to evaluating app-based mathematical interventions as it integrates several factors at different levels of the system. To further enhance its development, future research needs to consider the associations between app design and content (Grant et al. 2012) and children's learning outcomes with different software apps. This would help inform the design of app-based mathematics instruction to maximize children's learning opportunities and enhance the evidence base in this field.

General discussion

This study reports a novel, two-phased, systematic and pragmatic methodological approach for understanding the relationship between variation in the implementation of a maths app intervention and its impact on associated learning outcomes. This mixed-methods approach has potential to be applied to the implementation evaluation of other educational interventions and to set generic standards for this type of research, particularly at the early stages of a trialling an intervention. Before scaling, this method of statistically combining implementation process and impact evaluations can help identify which implementation factors are at play and most important for achieving the intended outcomes. This is a vital step for understanding how and why an intervention is successful and can help identify training needs in the development of further high stakes efficacy and effectiveness trials (Green et al. 2019). It could also contribute important insights to potential null findings and support educational research to be more informative (Lortie-Forgues and Inglis 2019). Overall, this new analytical method can add valuable contextual insights to quantitatively focused RCTs and thus enhance the epistemological ecosystem of educational enquiry (Thomas 2016; Shaffer 2011).

For other interventions where implementation is not commonly studied, the inductive approach utilized in phase I offers the opportunity to identify relevant themes that may not be currently available from previous research (Morse 1991). Furthermore, its qualitative nature affords a more in-depth insight into experiences of implementation and individual school contexts compared to more structured approaches, such as the implementation index used in previous research (e.g. Askell-Williams et al. 2013). In phase II, the structured judgements approach combined with the determinant theoretical framework enables causal inferences between intervention implementation and outcome to be established. This allows potential barriers and enablers to be identified (Nilsen 2015). In the context of the maths app intervention illustrated in the current study, 'established routine' was highlighted as a significant enabler. In the application to other intervention studies, this is a significant methodological advance for developing an understanding of the mechanisms underpinning educational interventions (Peterson 2016). Future studies have the potential to apply this approach incorporating other aspects of the multi-level, determinant framework (Bronfenbrenner 1979), such as examining potential barriers and enablers at the macro or individual level. Future studies could also include other dimensions of implementation (Durlak and DuPre 2008; Forman et al. 2009) to meet the needs of particular research questions.

It is important to acknowledge that the opportunity for implementation adaptations in the maths app intervention was relatively minimal compared to other more complex educational interventions. For example, in the current RCT curriculum content in the apps was fixed (Outhwaite et al. 2018), therefore adaptations ranging from minor, surface level changes, such as changing a cultural reference to appeal to the specific audience, to more substantial, deep changes, such as removing core components of the intervention (Moore, Bumbarger, and Cooper 2013), were not possible. In this study, adaptations focused on logistical delivery of the maths apps in individual classrooms and so may be more suited to observational measurement. In comparison, complex whole-school interventions that combine multiple interacting components including the situating context (Moore, Bumbarger, and Cooper 2013) may be more challenging to capture in the proposed mixed-methods approach (Anders et al. 2017). Consequently, further research is needed to apply and refine this new methodological approach for other interventions where implementation demands and opportunity for adaptations are greater. Furthermore, additional research utilizing this approach in this app-based mathematical learning context is required to establish the reliability of current results. If shown to replicate and be successful with a range of interventions, this new analytical approach has potential to contribute to the range of methodological tools available in implementation process evaluations (Evans, Scourfield, and Murphy 2015) and add greater insight to educational evaluation research.

Conclusion

The current study presents a new methodological approach for evaluating the impact of educational intervention implementation on learning outcomes. This innovative, two-phased, mixed-methods analytical approach makes a significant contribution to enhancing the epistemological ecosystem of educational enquiry. In illustrating the application of this approach, the importance of a well-established classroom routine was highlighted when implementing a maths app intervention within formal education settings. This has significant implications for scaling this intervention within primary schools to optimize effectiveness.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Economic and Social Research Council [grant number ES/J500100/1].

References

- Anders, J. D., C. Brown, M. Ehren, T. Greany, R. Nelson, J. Heal, A. Groot, M. Sanders, and R. Allen. 2017. *Evaluation of Complex Whole-school Interventions: Methodological and Practical Considerations*. London: Education Endowment Foundation.
- Anders, Y., H. G. Rossbach, S. Weinert, S. Ebert, S. Kuger, S. Lehl, and J. von Maurice. 2012. "Home and Preschool Learning Environments and Their Relations to the Development of Early Numeracy Skills." *Early Childhood Research Quarterly* 27 (2): 231–244.
- Askill-Williams, H., K. L. Dix, M. J. Lawson, and P. T. Slee. 2013. "Quality of Implementation of a School Mental Health Initiative and Changes Over Time in Students' Social and Emotional Competencies." *School Effectiveness and School Improvement* 24 (3): 357–381.
- Beach, R., and D. O'Brien. 2015. *Using Apps for Learning Across the Curriculum. A Literacy-based Framework and Guide*. New York: Routledge.
- Berkel, C., A. M. Mauricio, E. Schoenfelder, and I. N. Sandler. 2011. "Putting the Pieces Together: An Integrated Model of Program Implementation." *Prevention Science* 12 (1): 23–33.
- Biesta, G. J. 2010. "Why 'What Works' Still Won't Work: From Evidence-based Education to Value-based Education." *Studies in Philosophy and Education* 29 (5): 491–503.
- Blackwell, C. K., A. R. Lauricella, and E. Wartella. 2014. "Factors Influencing Digital Technology Use in Early Childhood Education." *Computers & Education* 77: 82–90.
- Brantlinger, E., R. Jimenez, J. Klingner, M. Pugach, and V. Richardson. 2005. "Qualitative Studies in Special Education." *Exceptional Children* 71 (2): 195–207.
- Bronfenbrenner, U. 1979. *The Ecology of Human Development*. Massachusetts: Harvard University Press.
- Clarke, D. 2004. "Structured Judgement Methods – the Best of Both Worlds?" In *Mixing Methods in Psychology: The Integration of Qualitative and Quantitative Practice*, edited by Z. Todd, B. Nerlich, S. McKeown, and D. Clarke, 79–99. Hove: Psychology Press.
- Clarke, V., and V. Braun. 2014. "Thematic Analysis." In *Encyclopedia of Quality of Life and Well-being Research*, edited by A. C. Michalos, 6626–6628. Dordrecht: Springer.
- Cohen, L., L. Manion, and K. Morrison. 2007. *Research Methods in Education*. New York: Routledge.
- Connolly, P., A. Biggart, S. Miller, L. O'Hare, and A. Thurston. 2017. *Using Randomised Controlled Trials in Education*. London: Sage.
- Connolly, P., C. Keenan, and K. 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.
- Cook, B. G., and S. L. Odom. 2013. "Evidence-based Practices and Implementation Science in Special Education." *Exceptional Children* 79 (2): 135–144.
- Couse, L. J., and D. W. Chen. 2010. "A Tablet Computer for Young Children? Exploring its Viability for Early Childhood Education." *Journal of Research on Technology in Education* 43 (1): 75–96.
- Cragg, L., S. Keeble, S. Richardson, H. E. Roome, and C. Gilmore. 2017. "Direct and Indirect Influences of Executive Functions on Mathematics Achievement." *Cognition* 162: 12–26.
- Creswell, J. W., and D. Creswell. 2018. *Research Design. Qualitative, Quantitative, & Mixed Methods Approaches*. London: Sage.
- Creswell, J. W., R. Shope, V. L. Plano Clark, and D. O. Green. 2006. "How Interpretive Qualitative Research Extends Mixed Methods Research." *Research in the Schools* 13 (1): 1–11.
- Domitrovich, C. E., S. D. Gest, D. Jones, S. Gill, and R. M. S. DeRousie. 2010. "Implementation Quality: Lessons Learned in the Context of the Head Start REDI Trial." *Early Childhood Research Quarterly* 25 (3): 284–298.
- Durlak, J. A., and E. P. DuPre. 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* 4: 327–350.
- Elliott, J. 2001. "Making Evidence-based Practice Educational." *British Educational Research Journal* 27 (5): 555–574.
- Elliott, D. S., and S. Mihalic. 2004. "Issues in Disseminating and Replicating Effective Prevention Programs." *Prevention Science* 5 (1): 47–53.
- Evans, R., J. Scourfield, and S. Murphy. 2015. "Pragmatic, Formative Process Evaluations of Complex Interventions and Why We Need More of Them." *Journal of Epidemiology and Community Health* 69: 925–926.
- Forman, S. G., S. S. Olin, K. E. Hoagwood, M. Crowe, and N. Saka. 2009. "Evidence-based Interventions in Schools: Developers' Views of Implementation Barriers and Facilitators." *School Mental Health* 1 (1): 26–36.
- Gambaro, L., K. Stewart, and J. Waldfogel. 2015. "A Question of Quality: Do Children from Disadvantaged Backgrounds Receive Lower Quality Early Childhood Education and Care?" *British Educational Research Journal* 41 (4): 553–574.
- Gathercole, S. E., and T. P. Alloway. 2006. "Practitioner Review: Short-term and Working Memory Impairments in Neurodevelopmental Disorders: Diagnosis and Remedial Support." *Journal of Child Psychology and Psychiatry* 47 (1): 4–15.
- Grant, A., E. Wood, A. Gottardo, M. A. Evans, L. Phillips, and R. Savage. 2012. "Assessing the Content and Quality of Commercially Available Reading Software Programs: Do They Have the Fundamental Structures to Promote the Development of Early Reading Skills in Children?" *NHSA Dialog* 15 (4): 319–342.

- Gray, P. 2015. *Free to Learn: Why Unleashing the Instinct to Play Will Make our Children Happier, More Self-reliant, and Better Students for Life*. New York: Basic Books.
- Green, C. S., D. Bavelier, A. F. Kramer, S. Vinogradov, U. Ansorge, K. K. Ball, U. Bingel, et al. 2019. "Improving Methodological Standards in Behavioral Interventions for Cognitive Enhancement." *Journal of Cognitive Enhancement* 3 (1): 2–29.
- Hammersley, M. 2007. *Educational Research and Evidence-based Practice*. London: Sage.
- Haynes, P., B. Chambers, and J. Haslam. 2016. "Reassessing RCTs as the 'Gold Standard': Synergy Not Separatism in Evaluation Designs." *International Journal of Research & Method in Education* 39 (3): 287–298.
- Hansen, W. B. 2014. "Measuring Fidelity." In *Defining Prevention Science*, edited by Z. Sloboda and H. Petras, 335–359. New York: Springer.
- Hansen, W. B., M. M. Pankratz, L. Dusenbury, S. M. Giles, D. C. Bishop, J. Albritton, L. P. Albritton, and J. Strack. 2013. "Styles of Adaptation: The Impact of Frequency and Valence of Adaptation on Preventing Substance Use." *Health Education* 113 (4): 345–363.
- Haynes, L., O. Service, B. Goldacre, and D. Torgerson. 2012. *Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials*. London: Cabinet Office.
- Herodotou, C. 2018. "Young Children and Tablets: A Systematic Review of Effects on Learning and Development." *Journal of Computer Assisted Learning* 34 (1): 1–9.
- Hilton, A. 2016. "Engaging Primary School Students in Mathematics: Can iPads Make a Difference?" *International Journal of Science and Mathematics Education*. doi:10.1007/s10763-016-9771-5.
- Hirsh-Pasek, K., J. M. Zosh, R. M. Golinkoff, J. H. Gray, M. B. Robb, and J. Kaufman. 2015. "Putting Education in 'Educational' Apps: Lessons from the Science of Learning." *Psychology Science* 16 (1): 3–34.
- 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, 150–163. Maidenhead: Open University Press.
- Hoffmann, T. C., P. P. Glasziou, I. Boutron, R. Milne, R. Perera, D. Moher, D. G. Altman, et al. 2014. "Better Reporting of Interventions: Template for Intervention Description and Replication (TIDieR) Checklist and Guide." *BMJ* 348: g1687.
- Humphrey, N. 2013. *Social and Emotional Learning: A Critical Appraisal*. London: Sage.
- Humphrey, N., A. Lendrum, E. Ashworth, K. Frearson, R. Buck, and K. Kerr. 2016. *Implementation and Process Evaluation (IPE) for Interventions in Educational Settings: A Synthesis of the Literature*. London: Education Endowment Foundation.
- Irwin, D. M., and M. M. Bushnell. 1980. *Observational Strategies for Child Study*. New York: Holt, Rinehart, and Winston.
- Kalaycioğlu, D. B. 2015. "The Influence of Socioeconomic Status, Self-efficacy, and Anxiety on Mathematics Achievement in England, Greece, Hong Kong, the Netherlands, Turkey, and the USA." *Educational Sciences: Theory and Practice* 15 (5): 1391–1401.
- Kirschner, P. A., J. Sweller, and R. E. Clark. 2006. "Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-based, Experiential, and Inquiry-based Teaching." *Educational Psychologist* 41 (2): 75–86.
- Kucian, K., U. Grond, S. Rotzer, B. Henzi, C. Schönmann, F. Plangger, M. Gälli, E. Martin, and M. von Aster. 2011. "Mental Number Line Training in Children with Developmental Dyscalculia." *Neuroimage* 57 (3): 782–795.
- Kucirkova, N. 2014. "iPads in Early Education: Separating Assumptions and Evidence." *Frontiers in Psychology* 5: 715. doi:10.3389/fpsyg.2014.00715.
- Langley, G. J., R. D. Moen, K. M. Nolan, T. W. Nolan, C. L. Norman, and L. P. Provost. 2009. *The Improvement Guide: A Practical Approach to Enhancing Organizational Performance*. San Francisco: Jossey-Bass.
- Lendrum, A., N. Humphrey, and M. Greenberg. 2016. "Implementing for Success in School-based Mental Health Promotion: The Role of Quality in Resolving the Tension Between Fidelity and Adaptation." In *Mental Health and Wellbeing Through Schools: The Way Forward*, edited by R. Shute and P. Slee, 53–63. London: Taylor and Francis.
- Lortie-Forgues, H., and M. Inglis. 2019. "Rigorous Large-scale Educational RCTs are Often Uninformative: Should We Be Concerned?" *Educational Researcher* 48 (3): 158–166.
- Miles, M. B., and A. M. Huberman. 1994. *Qualitative Data Analysis: An Expanded Sourcebook*. London: Sage.
- Moore, J. E., B. K. Bumbarger, and B. R. Cooper. 2013. "Examining Adaptations of Evidence-based Programs in Natural Contexts." *The Journal of Primary Prevention* 34 (3): 147–161.
- Morse, J. M. 1991. "Approaches to Qualitative-quantitative Methodological Triangulation." *Nursing Research* 40 (2): 120–123.
- Naylor, P. J., L. Nettlefold, D. Race, C. Hoy, M. C. Ashe, J. W. Higgins, and H. A. McKay. 2015. "Implementation of School Based Physical Activity Interventions: A Systematic Review." *Preventive Medicine* 72: 95–115.
- Nilsen, P. 2015. "Making Sense of Implementation Theories, Models and Frameworks." *Implementation Science* 10 (53): 1–13.
- Oakley, A., V. Strange, C. Bonell, E. Allen, and J. Stephenson. 2006. "Process Evaluation in Randomised Controlled Trials of Complex Interventions." *BMJ* 332 (7538): 413–416.
- O'Donnell, C. L. 2008. "Defining, Conceptualizing, and Measuring Fidelity of Implementation and its Relationship to Outcomes in K–12 Curriculum Intervention Research." *Review of Educational Research* 78 (1): 33–84.
- Ofsted. 2013. *Annual Report 2012/2013: East Midlands Regional Report*. London: Ofsted.
- Ofsted. n.d. *Inspection Reports*. Retrieved from <https://reports.beta.ofsted.gov.uk> for each participating school.

- Outhwaite, L. A., M. Faulder, A. Gulliford, and N. J. Pitchford. 2018. "Raising Early Achievement in Math with Interactive Apps: A Randomized Control Trial." *Journal of Educational Psychology* 111 (2): 284–298.
- Outhwaite, L. A., A. Gulliford, and N. J. Pitchford. 2017. "Closing the Gap: Efficacy of a Tablet Intervention to Support the Development of Early Mathematical Skills in UK Primary School Children." *Computers & Education* 108: 43–58.
- Pawson, R., and N. Tilley. 1997. "Realistic Evaluation." In *Encyclopedia of Evaluation*, edited by S. Matthieson, 359–367. London: Sage.
- Peterson, A. 2016. "Getting 'What Works' Working: Building Blocks for the Integration of Experimental and Improvement Science." *International Journal of Research & Method in Education* 39 (3): 299–313.
- Pitchford, N. J. 2015. "Development of Early Mathematical Skills with a Tablet Intervention: A Randomized Control Trial in Malawi." *Frontiers in Psychology* 6: 485. doi:10.3389/fpsyg.2015.00485.
- Punch, K. F. 2013. *Introduction to Social Research: Quantitative and Qualitative Approaches*. London: Sage.
- Robson, C., and K. McCartan. 2011. *Real World Research*. Chichester: Wiley.
- Rogers, E. M. 2003. *The Diffusion of Innovations*. New York: Free Press.
- Saldaña, J. 2015. *The Coding Manual for Qualitative Researchers*. London: Sage.
- Schagen, S., S. Blenkinsop, I. Schagen, E. Scott, M. Eggers, I. Warwick, E. Chase, and P. Aggleton. 2005. "Evaluating the Impact of the National Healthy School Standard: Using National Datasets." *Health Education Research* 20 (6): 688–696.
- Shaffer, P. 2011. "Against Excessive Rhetoric in Impact Assessment: Overstating the Case for Randomised Controlled Experiments." *Journal of Development Studies* 47 (11): 1619–1635.
- Slavin, R. 2012. "Foreward." In *Handbook of Implementation Science for Psychology in Education*, edited by B. Kelly, and D. F. Perkins, xv. Cambridge: Cambridge University Press.
- Strand, S., L. Malmberg, and J. Hall. 2015. *English as an Additional Language (EAL) and Educational Achievement in England: An Analysis of the National Pupil Database*. London: Education Endowment Foundation.
- Tashakkori, A., and C. Teddlie. 2010. "Putting the Human Back in 'Human Research Methodology': The Researcher in Mixed Methods Research." *Journal of Mixed Methods Research* 4 (4): 271–277.
- Thomas, G. 2016. "After the Gold Rush: Questioning the 'Gold Standard' and Reappraising the Status of Experiment and Randomized Controlled Trials in Education." *Harvard Educational Review* 86 (3): 390–411.
- Thomas, G., and R. Pring. 2004. *Evidence-based Practice in Education*. Maidenhead: Open University Press.
- Torgerson, D. J., and C. J. Torgerson. 2003. "Avoiding Bias in Randomised Controlled Trials in Educational Research." *British Journal of Educational Studies* 51 (1): 36–45.
- Vignoles, A., J. Jerrim, and R. Cowan. 2015. *Mathematics Mastery: Primary Evaluation Report*. London: Education Endowment Foundation.
- Wyse, D., and C. Torgerson. 2017. "Experimental Trials and 'What Works?' In Education: The Case of Grammar for Writing." *British Educational Research Journal* 43 (6): 1019–1047.