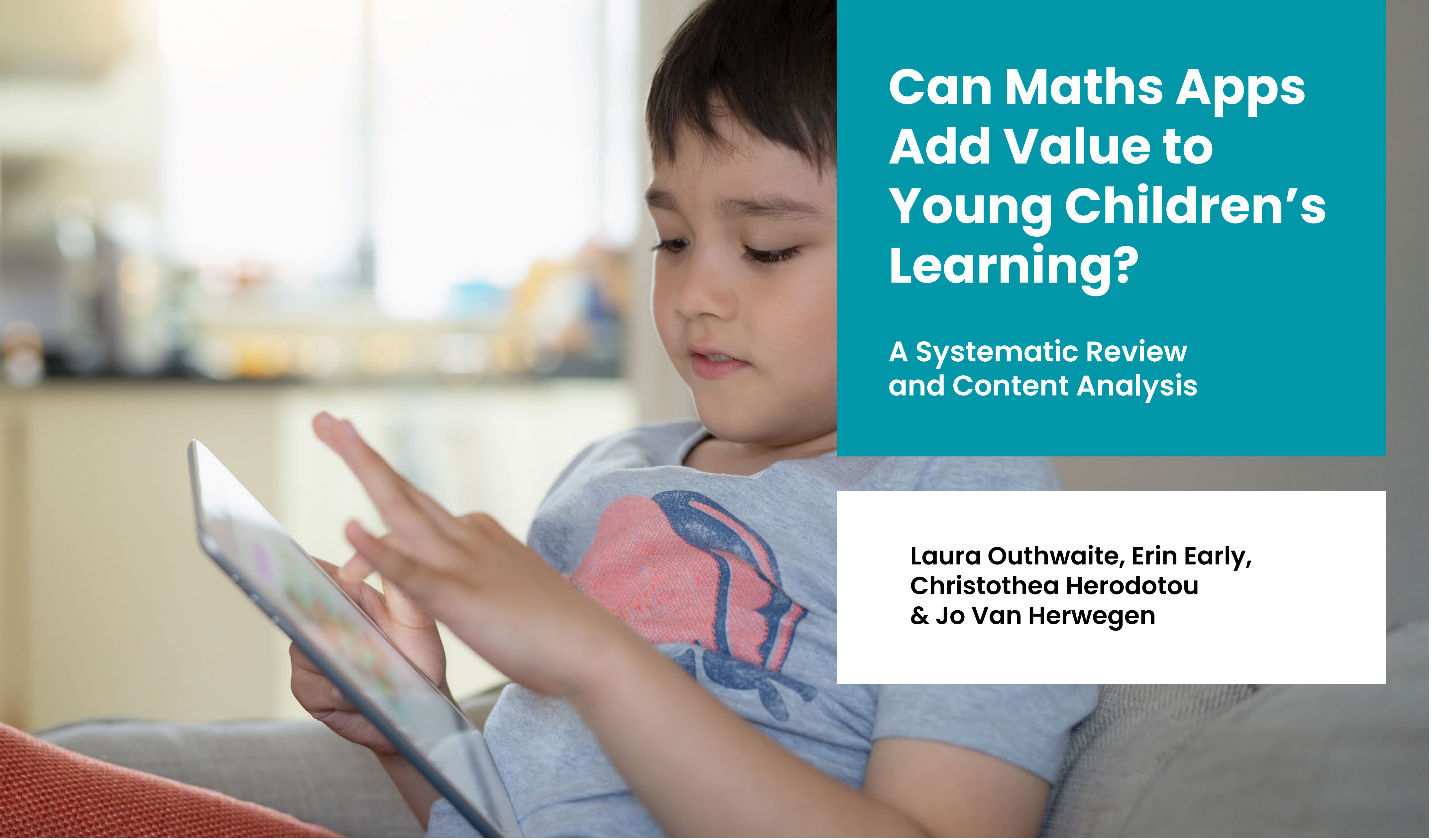


Can Maths Apps Add Value to Young Children's Learning?

A Systematic Review and Content Analysis

Laura Outhwaite, Erin Early,
Christothea Herodotou
& Jo Van Herwegen

IOE – FACULTY OF
EDUCATION AND SOCIETY

The Open
University

Nuffield
Foundation

Table of Contents

About the Authors

Dr Laura Outhwaite is a Senior Research Fellow in the Centre for Education Policy and Equalising Opportunities at IOE, UCL's Faculty of Education and Society.

Dr Erin Early is a Research Fellow in the Centre for Education Policy and Equalising Opportunities at IOE.

Dr Christothea Herodotou is a Senior Lecturer in the Institute of Educational Technology at the Open University.

Dr Jo Van Herwegen is an Associate Professor in the Psychology and Human Development department at IOE. She is also a member of the Centre for Education Policy and Equalising Opportunities.

Advisory Board

Prof Alison Clark-Wilson is a Professorial Research Fellow at the UCL Knowledge Lab, IOE.

Prof Natalia Kucirkova is a Professor of Reading and Early Childhood Development at the University of Stavanger.

Ms Ruth Maisey is a Programme Head for Education at the Nuffield Foundation.

Dr Alison O'Mara-Eves is an Associate Professor in the Evidence for Policy and Practice Information (EPPI) Co-ordinating Centre at IOE.

Prof Victoria Simms is a Professor of Developmental Psychology in the School of Psychology at Ulster University.

Ms Caroline Wootton is a specialist mathematics teacher for primary school children struggling to learn mathematics, including children with special educational needs.

Ms Lucy Williams is a Senior Early Years Adviser for the UK Impact Team at Save the Children.

Acknowledgements

The Nuffield Foundation is an independent charitable trust with a mission to advance social well-being. It funds research that informs social policy, primarily in Education, Welfare, and Justice. It also funds student programmes that provide opportunities for young people to develop skills in quantitative and scientific methods. The Nuffield Foundation is the founder and co-funder of the Nuffield Council on Bioethics, the Ada Lovelace Institute and the Nuffield Family Justice Observatory. The Foundation has funded this project, but the views expressed are those of the authors and not necessarily the Foundation.

Visit **www.nuffieldfoundation.org**

Executive Summary

Executive Summary

Educational maths applications (apps) are an emerging trend in children's learning environments aiming to raise their mathematical attainment. However, with over 200,000 educational apps available within the App Store (Apple, 2014), deciding which apps to use poses a significant challenge to teachers, parents, and policy makers. The current study aimed to advance our understandings of whether and how educational maths apps can support children's learning, as well as outline gaps in current research evidence and practice. In doing so, the current study included:

- **A systematic review** (Part 1) to synthesise the current evidence on educational maths apps for young children in the first three years of compulsory school (e.g., ages 4-7 years in England; ages 5-8 years in the USA).
- **A content analysis** (Part 2) to examine the content and design features of different educational maths apps and how they may support children's learning.

What do we mean by educational maths apps?

Educational maths apps are interactive software primarily used on a hand-held touch-screen tablet or smartphone device with the aim of supporting learning outcomes. There are different types of educational apps, which have different aims and design features (Kay & Kwak, 2018):

- **Practice-based apps** (e.g., *onebillion Maths 3-5*) are designed to support the acquisition of learning content, such as mathematical facts and concepts through targeted practice. They are primarily designed to be used by children individually and are self-paced. In practice-based apps, the child is the consumer of the learning content.
- **Game-based apps** (e.g., *Slice Fractions*) are like practice-based apps, with the addition that the learning content is embedded within a broader immersive player narrative.
- **Constructive apps** (e.g., *Montessori Numbers for Kids*) are designed to encourage the exploration and active manipulation of mathematical ideas and concepts. With constructive apps, the child is still the consumer of the learning content.
- **Productive apps** (e.g., *Quizlet Plus*) are designed to support children to produce their content, for example, to present their own ideas on a particular maths topic. As such, the child is the creator of the learning content.
- **Parent-based apps** (e.g., *Bedtime Math*), are primarily designed to be used by parents or caregivers and are designed to encourage offline interactions and learning opportunities with children.

Part 1: Systematic Review

What studies were included?

The systematic review included studies that assessed the impact of educational maths apps used at home or at school for children in the first three years of compulsory school on mathematical attainment and other outcomes, including cognitive development, enjoyment, and motivation. Fifty studies were identified that included both quantitative and qualitative findings.

What were the key findings relating to evaluation methods, populations, and outcomes?

Evaluation methods

- The most common research designs adopted experimental methods (n = 33), with 20 randomised control trials and 13 quasi-experimental designs.
- Across these 33 experimental studies, 21 showed positive results for mathematical learning outcomes in favour of the maths apps, compared to a range of control groups.
- Eight experimental studies showed mixed results, and four studies found no difference in learning outcomes between the group using the maths apps and the control groups.

Populations

- Most studies focused on the use of educational maths apps with typically developing children of all attainment levels (n = 43).
- Initial evidence suggested that educational maths apps can also support children identified as underachieving in mathematics (n = 8) and children with special educational needs and disabilities (SEND) (n =3).

Outcomes

- The included studies mostly focused on mathematical learning outcomes immediately following the intervention period (n = 47).
- Only five studies included a follow-up (or delayed) assessment of children's mathematical attainment, ranging from one month to two years with mixed results.
- Thirteen studies also included a range of non-attainment outcome measures. Results mostly showed children enjoyed using the maths apps, but mixed results were found in terms of children's intrinsic motivation, flow experience, and preference towards app-based learning, compared to other maths activities.

Part 2: Content Analysis

The 50 studies included in the systematic review predominately reported greater learning outcomes for children using the evaluated maths apps compared to a range of control conditions. However, little research has considered what the active ingredients (i.e., mechanisms), or combination of ingredients, of successful maths apps are, and how they link to current theories of mathematical development and learning science. To examine how different maths apps work, the underpinning pedagogy, including the mathematical content and app design features, needs to be examined (Griffith et al., 2020) and linked to the observed learning outcomes (Outhwaite et al., 2019).

This research sought to do so in three steps, which formed part of the new framework for educational maths apps:

a. A content analysis of the mathematical content and app design features was conducted with the educational maths apps identified in the systematic review (Part 1). This content analysis drew on existing frameworks on the educational value of apps and developed a new framework that adopted a combined top-down (deductive) and bottom-up (inductive) approach.

b. A qualitative comparative analysis (QCA) was conducted and identified the specific app design features, or combination of features, that were sufficient to support children's learning outcomes with educational maths apps.

c. The coding framework developed for the content analysis was then also applied to the Top 25 most popular commercial educational maths apps from the iOS Apple App and Google Play Stores.

What apps were included?

a. To be eligible for the content analysis, apps identified through the systematic review (Part 1) needed to meet the following criteria:

- Apps were the individual focus of an intervention study with maths attainment as the primary outcome measure.
- If multiple apps were included within a single study, the reported results were disaggregated for each app.
- Apps were commercially available and accessible for download in the UK.

23 maths apps met these criteria.

b. To be eligible for the QCA, apps identified through the systematic review (Part 1) needed to meet the above criteria. In addition, apps also needed to be evaluated in studies with sufficiently reported data to calculate the within-subject effect sizes (Cohen's *d*). 8 apps met these criteria.

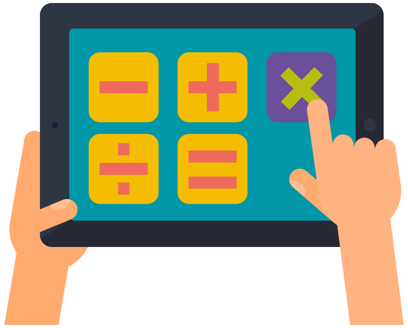
c. The most popular commercial educational maths apps for children aged 0-5 years were selected from the Apple and Google Play Stores in August 2021. 25 apps were considered.

What were the key findings relating to types of apps evaluated, their mathematical content, and design features?

Type of apps

Type of app	Number of apps
Practice-based	15
Game-based	4
Constructive	2
Productive	1
Parent-based	1
Total number of apps	*23*

Practice-based maths were the most common. They are designed to support the acquisition of learning content through targeted practice.



Mathematical content

123

The most common area of maths covered in the maths apps was understanding number representations and relationships.

83%

of the maths apps covered number representation with Arabic digits, verbal, and/or written number word recognition.

78%

of the maths apps taught one-to-one correspondence. This is the understanding that each item in a set can only be counted once.

61%

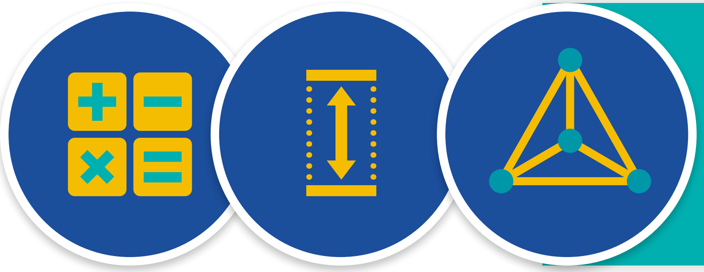
of the maths apps covered cardinality. This is the understanding that the last number counted in a set represents the overall total.

52%

of the maths apps targeted single digit addition and subtraction.

57%

of the maths apps included shape, patterns, and measurement skills.



More advanced maths skills, including fractions, place value, multiplication, and sequencing events were the least common mathematical skills covered across the maths apps.

Design features: Feedback

35% of the maths apps provided motivational feedback, such as “You did it!” or “Great job!”.

56% of the maths apps provided motivational and explanatory feedback. Explanatory feedback provides children with an explanation of why their answer is correct or incorrect.

9% of the maths apps gave no feedback.

Design features: Levelling

52% of the maths apps provided a suggested, but not enforced sequence of learning content. This is known as participatory, free form levelling.

22% of the maths apps provided levelling that was adaptive of learning content in response to a child's performance when using the app/ This is known as programmatic, dynamic levelling.

17% of the maths apps provided learning content that was tailored to a child based on an initial attainment assessment or learning content that was pre-selected by an adult but did not adapt based on their in-app behaviour. This is known as programmatic, static levelling.

9% of the maths apps gave no levelling.

What were the key findings from the QCA?

- Children's learning outcomes with maths apps were maximised when the apps provided a scaffolded and personalised learning journey (programmatic levelling) and explained why their answer was right or wrong (explanatory feedback), as well as giving praise, such as "Great job!" (motivational feedback).

Most popular commercial maths apps

- The new framework developed in the content analysis in Part 2 (type of app, mathematical content, and app design features) was then applied to the Top 25 most popular commercial educational maths apps available on the Apple and Google Play Stores for children aged 5 years.

- Most commercial apps were also identified as practice-based and primarily targeted number and counting skills.

- However, only one app identified in the Top 25 has been empirically evaluated (*onebillion Maths 4-6*).

- Six of the most popular commercial educational apps identified as maths on the search terms did not explicitly include any maths content.

Recommendations

For researchers

Future studies in this area need to:

- Consider the skills targeted by the apps (near-transfer), as well as the generalisation to broader mathematical skills and related non-attainment outcomes (far-transfer), including within a longer-term follow-up after the intervention study has finished.

- Evaluate the use of educational maths apps with young children identified as underachieving in mathematics, and these children should be reliably identified in ways that do not threaten the internal validity of the findings.

- Evaluate the use of educational maths apps with children with SEND, such as dyscalculia, Down syndrome, and other learning difficulties.

- Further examine the role of children's language skills and their age when using different types of maths apps.

- Rigorously capture children's time on task when using the apps to examine how this is associated with their immediate and sustained learning outcomes.

- Make use of innovative methods of data collection to capture in-app data on learner analytics and enable further work with hard-to-reach populations.

- Examine the use educational maths apps in different educational, cultural, and economic contexts in cross-cultural studies.

- Evaluate educational maths apps in different settings, particularly the home learning environment, for addressing educational challenges, including those raised during the Covid-19 pandemic.

Overall, research in this area would benefit from improved reporting standards, whereby sufficient information is reported to ensure studies are rigorously conducted and evaluated. This will support future synthesises and increase the potential to impact on policy and practice.

For parents and teachers

When deciding if and which educational maths apps to use with children, teachers and parents should consider three key things:

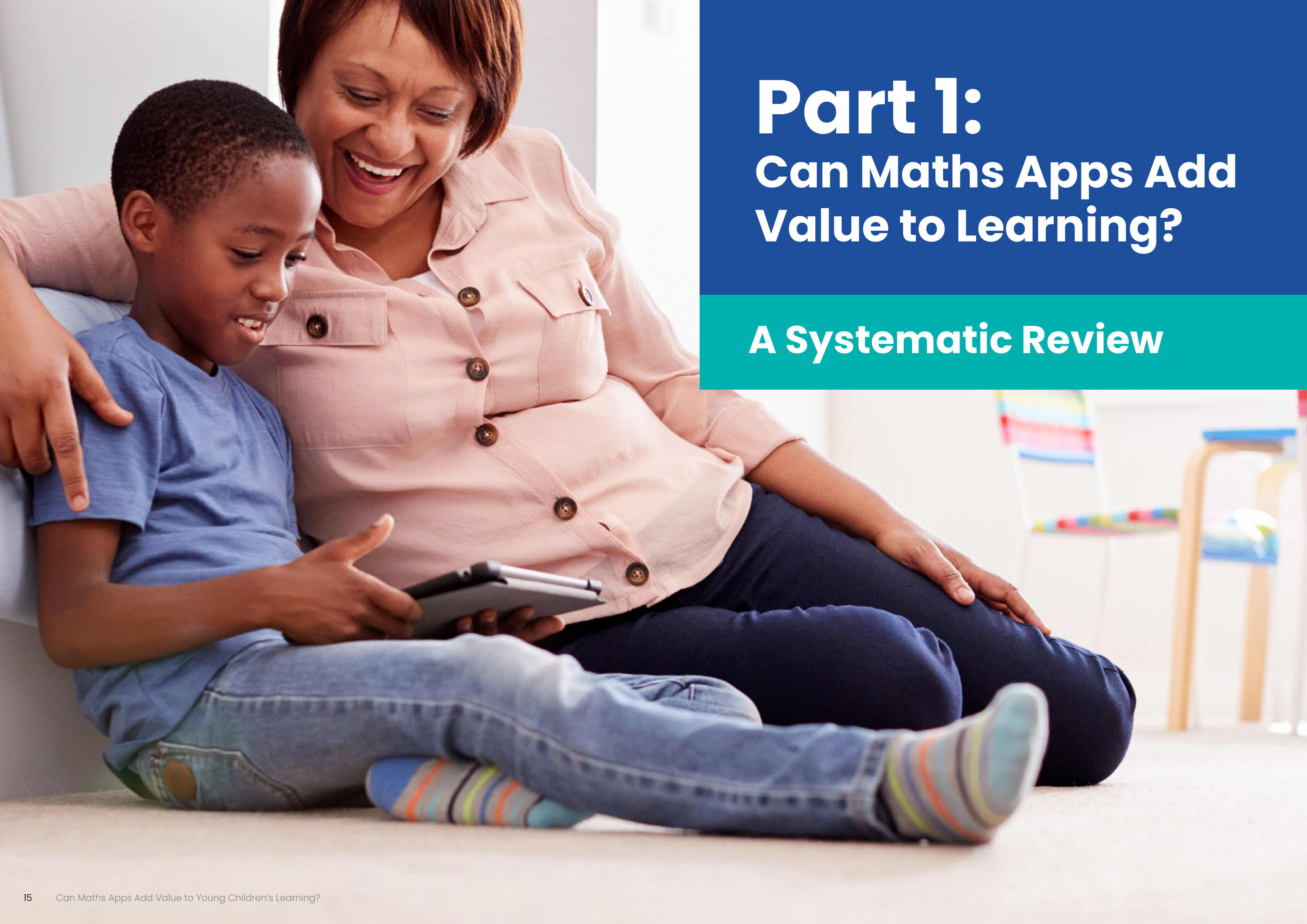
- When trying the apps for the first time, consider whether their child can meaningfully interact with the chosen maths apps, particularly based on their existing mathematical and language skills?

- How can the chosen maths apps be implemented in their own classroom or home? For example, does the app include areas of maths that children need extra support with? What support may children need to use the apps effectively?

- Does the chosen app contain explanatory feedback, whereby the app explains why a response is correct or incorrect and programmatic levelling, whereby the learning content is scaffolded and personalised to the child?

For policy makers

- To ensure the best outcomes for all children, information about the educational content and quality of the app design needs to be readily available (e.g., on the Apple and Google Play Stores) and carefully considered.
- Considerations should also be given as to how the apps can be accessed, selected, and used by schools and families at home if they choose. As such, parents and teachers also need access to appropriate support, for example training and guidance on best practices informed by the users themselves.



Part 1:

Can Maths Apps Add Value to Learning?

A Systematic Review

The Need for the Systematic Review

In the UK, 20% of 5-year-old children do not achieve the expected attainment for mathematics (Department for Education [DfE], 2017). Developing strong early mathematical skills is vital for children's later educational success (Clements & Sarama, 2009) and their economic, health, and employment outcomes (Reyna et al., 2009). To address mathematical underachievement, evidence-based interventions are needed that successfully engage children from a young age (Jordan & Levine, 2009). An emerging trend aiming to benefit children's mathematical learning is the use of educational maths applications (apps) delivered on touch-screen tablets (e.g., iPads/ Androids) in school and at home (DfE, 2019). These apps are better suited for children over 4 years compared to younger children under 3 years (Herodotou, 2018).

In the UK, 94% of children aged 3-11 years have access to touch-screen tablet devices (Marsh et al., 2020) with parents of pre-school aged children most likely to download educational apps (Chaudron, 2015). 41% of teachers also reported that they use maths apps as supplementary teaching tools in early primary school (Vega & Robb, 2019). More broadly, UK schools spend over £900 million a year on educational technology (London & Partners, 2015) and as a rapidly growing sector, spurred by the Covid-19 pandemic and the increased need for digital learning environments, it is expected to be worth £3.4 billion this year (Education Technology, 2021).

Alongside this growth, interest and use of educational maths apps, there is also increasing concern regarding young children's screen time and the impact it has on early learning and development (American Academy of Pediatrics, 2016). Some recent studies have demonstrated the benefits of different maths apps for children (Outhwaite et al., 2018; Berkowitz et al., 2015). However, there are mixed results across different maths apps (Parks & Tortorelli, 2020). To examine how different educational maths apps work, the underpinning pedagogy, including app content and design features, needs to be examined (Outhwaite et al., 2019).

Previous Reviews

In a recent review, Herodotou (2018) included 19 experimental studies examining the use of educational apps (i.e., not just maths) with children aged 2-5 years. Five studies focused on mathematics and found positive benefits for 4-year-old children compared to younger 3-year-old children. Griffith et al. (2020) also identified 15 experimental studies that demonstrated interactive apps can support mathematical learning outcomes and called for more research examining app design features and large-scale evaluations of educational apps. In another recent systematic review, Simms et al. (2019) identified 11 studies that used computerised instructional programmes for children aged 4-11 years with mixed results. Although the review by Simms et al. (2019) takes a broader age range compared to other systematic reviews (Griffith et al., 2020; Herodotou, 2018), it does not focus explicitly on educational apps.

Overall, existing reviews have not comprehensively synthesised the current quantitative and qualitative research evidence to examine whether educational maths apps provide an effective form of instruction for mathematical and non-attainment outcomes with young children at the start of formal education. For example, qualitative studies, which were not included in previous reviews, provide useful insights into how the implementation of the maths app intervention can impact outcomes (e.g., Outhwaite et al., 2019) and how app features can support or hinder children's engagement and learning (e.g., Herodotou, 2021; Moyer-Packenham et al., 2016). Likewise, feasibility studies (e.g., Outhwaite et al., 2017), when successful, can support the development of further studies focusing on understanding the mechanisms and efficacy of the maths app interventions (Green et al., 2019).

Aims and Objectives of the Current Review

In response, the current review aimed to comprehensively synthesise the current evidence to capture existing knowledge and identify important gaps on whether maths apps can support learning, under what circumstances, and how maths apps may work to support learning, as well as for whom maths apps may benefit.

More specifically, this study systematically reviewed the quantitative and qualitative research literature that has examined the impact of educational maths apps used at school or at home. The review focused on mathematical learning outcomes (primary outcome measure) and considered other outcomes, including cognitive development, enjoyment, and motivation (secondary outcome measures). The review also focused on children in the first three years of compulsory school (e.g., Reception to Year 2 in England where children will be aged 4-7 years). Included studies were synthesised in a comprehensive thematic narrative synthesis with key findings and knowledge gaps identified (Snilstveit et al., 2016) to address the following four Research Questions (RQs):

RQ1: What type, frequency, and quality of evaluation studies have been conducted with educational maths apps?

RQ2: What lessons can be learnt from previous research examining educational maths apps, relating to methods, populations, and outcomes?

RQ3: What is the external validity of the current evidence?

RQ4: What gaps can be identified based on the current evidence?

Overall, this study provides a comprehensive overview of the current international research evidence on educational maths apps. This allows for a better understanding of 1) current research quality, 2) current and new methodological considerations and 3) where the gaps are in research. For academics, this generates a clearly articulated new research agenda at a time when app-based learning and educational technology more broadly are growing rapidly in popularity, usage, and investment. Similarly, for teachers, parents, and policymakers this develops insight

into "what works" and for whom in the context of educational maths apps and provides the grounding for evidence-based guidance.

Methods

The systematic literature search was conducted using detailed and pre-defined search terms, as well as inclusion and exclusion criteria developed using the PICO (Population, Intervention, Comparison, and Outcome) method to select relevant studies for the thematic narrative synthesis. The systematic review protocol was pre-registered on the Open Science Framework **(https://osf.io/pzkmh/)**.

Search strategy

A broad search of the literature was conducted in January 2021 using search terms focused on the population, intervention, and outcome components of the PICO statement (see Table 1). As touch-screen tablets were first introduced in 2010, the search focused on studies published between 2010- 2021.

Table 1: Search strategy used in the systematic review.

PICO	Search String
Population	("early years" OR preschool* OR kindergart* OR "primary school" OR "elementary school" OR "young children" OR "young pupils" OR "young students" OR child* OR pupils OR students) AND
Intervention	"educational app" OR "interactive app" OR tablet OR "tablet technology" OR "iPad app" OR "android app" OR "math* app") AND
Outcome	(math* OR number* OR "number sense" OR arithmetic* OR measurement OR geomet* OR shape)

Electronic web searches of published and grey literature were completed using the databases and platforms outlined in Table 2.

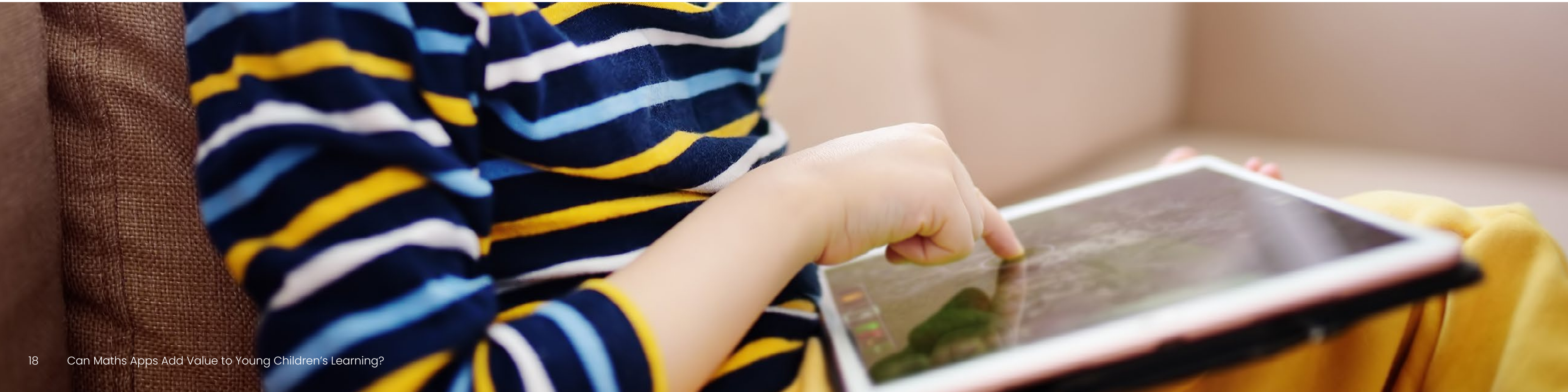


Table 2: Databases searched using the listed platforms.

Database	Platform
PsycINFO	Ovid
Educational Resource Information Centre (ERIC)	Ovid
Medline	Ovid
Scopus	Ovid
Science Citation Index (SC) Expanded	Web of Science
Social Science Citation Index (SSCI)	Web of Science
Arts & Humanities Citation Index (A&HCI)	Web of Science
Emerging Sources Citation Index	Web of Science
PubMed	https://pubmed.ncbi.nlm.nih.gov/
British Education Index (BEI)	UCL Explore
Australian Education Index (AEI)	UCL Explore
Dissertations and theses	ProQuest (abstract only)
Open Science Framework preprints	https://osf.io/preprints/
PsyArXiv preprints	https://psyarxiv.com/
Nuffield Foundation Research Reports	https://www.nuffieldfoundation.org/research (search string limited to "math*")
Education Endowment Foundation completed projects	https://educationendowmentfoundation.org.uk/ (search string limited to "math*")

In addition, a backwards citation search of three recent and relevant systematic reviews (Griffith et al., 2020; Herodotou, 2018; Simms et al., 2019). was conducted. These existing reviews synthesise evidence on mathematical interventions and educational apps in general (i.e., not specific to maths), and thus were selected based on the similarities with the current review. A prospective forward citation search of included studies (n = 45) was also conducted in June 2021.

Inclusion criteria

For inclusion in the systematic review, studies needed to meet the following eligibility criteria, which was developed based on the PICO Framework.

Population: Studies needed to include children in the first three years of compulsory school. This age group was selected based on the emerging trend in the use of educational apps at the start of formal education (DfE, 2019) and evidence suggesting apps are more beneficial for children over 4 years (Herodotou, 2018).

However, as different countries have different age ranges for the first three years of compulsory education (e.g., children start preschool age 6 in Finland) or they may follow an ability-based system, rather than an age-based system (e.g., Malawi), the ages of children varied across the included studies. In England, the first three years of compulsory school refers to Reception to Year 2, where children are aged 4-7 years. In the USA, this refers to Kindergarten to Grade 2, where children are aged 5-8 years. In some European countries, such as Finland, this refers to preschool to Grade 2, where children are aged 6-9 years. Whereas, in Malawi, this refers to Standard 1-Standard 3, where children typically start school aged 6 years. However, while primary education is free in Malawi, it is not compulsory, and many children start school at different ages. The repetition of school years is also common. This means a child may be in an academic year that is not typically aligned with their chronological age.

If studies included children within the specified school years, as well as older children, only data relating to the year groups of interest were extracted. Corresponding authors were contacted where necessary. Studies that focused on typically developing children, children underachieving in mathematics and children with special educational needs were eligible for inclusion.

Intervention: Included studies needed to evaluate a named downloadable maths app(s) and not just a particular feature of an app. Apps needed to be commercially available, or researcher developed. For the purposes of this review, an educational app was defined as interactive software that is primarily designed to be used on a hand-held touch-screen tablet or smartphone devices. It did not include computer-assisted instruction software e.g., web-based programmes that can only be used on computers but included apps that were connected to external/physical manipulatives.

Studies also needed to be available in English and included details on the duration (i.e., number of weeks) and intensity of the intervention

implementation (i.e., number of sessions per week and length of sessions). Data on intervention duration and intensity was multiplied to give a 'total time on task'.

Comparison: To reflect the staged development of maths app evaluations, the review did not place any restrictions on the study design used for the thematic narrative synthesis. Included studies may or may not have included a control group comparison in a quasi-experimental (i.e., utilising pre-existing groups or within-subject design) or randomised control trial (RCT) design. No restriction was placed on the type of control group (e.g., active control or business-as-usual) or number of intervention groups.

Outcome: The primary unit of analysis was mathematical learning outcomes in response to a specific maths app(s) used at school or at home immediately following the intervention period. Other outcomes were also considered including cognitive development, enjoyment, and motivation (secondary outcomes). No restriction was placed on whether outcome measures were researcher developed or standardised. However, this information was extracted from the included studies and its impact on the conclusions of the review were explored in the thematic narrative synthesis.

Study screening

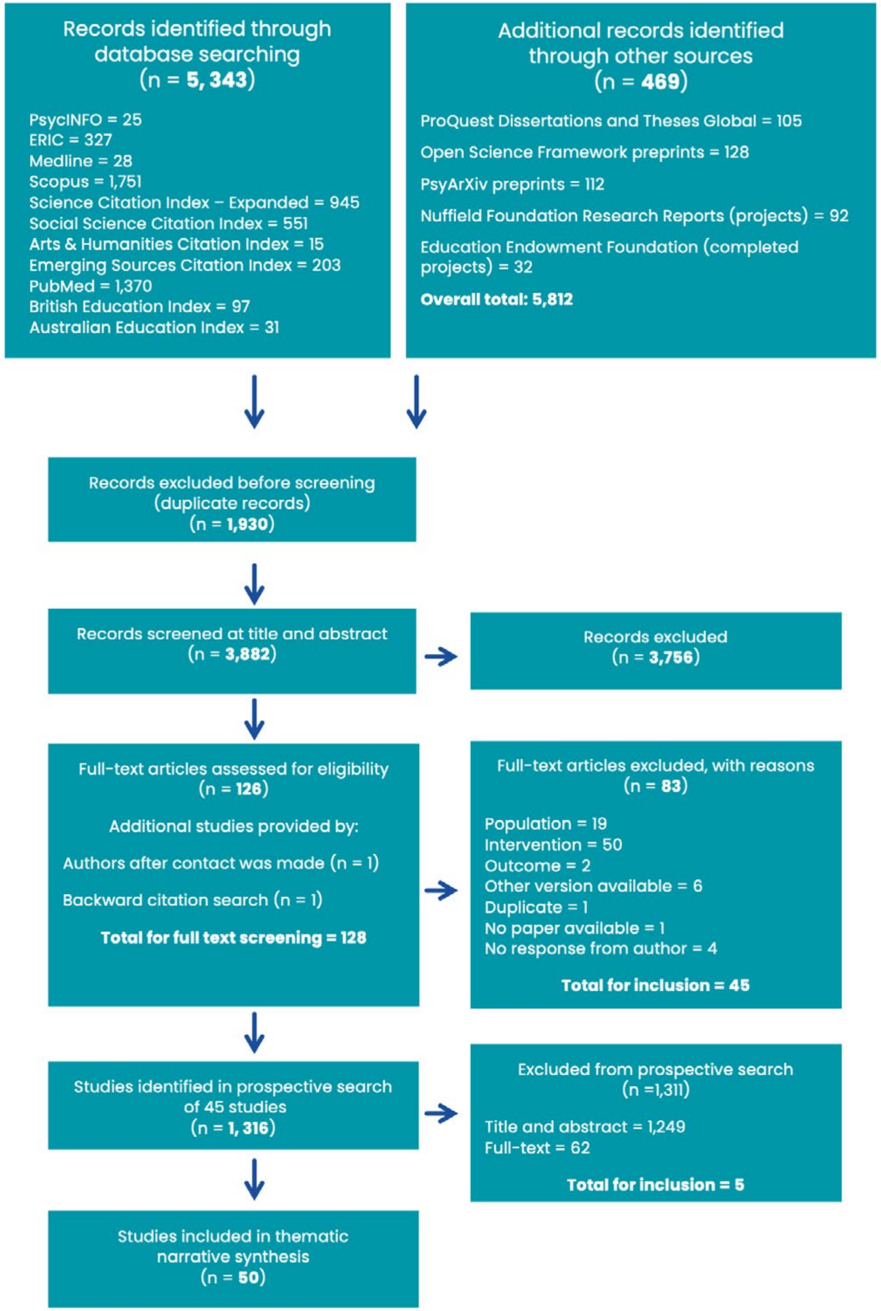
As outlined in the PRISMA Flow Diagram (Page et al., 2021; see Figure 1), the database searches returned 5,812 records. The titles and abstracts of the records returned in the database searches were imported to the web-based software EPPI-Reviewer where 1,930 duplicate records were identified and removed, resulting in 3,882 unique records. Unique records were screened based on the inclusion and exclusion criteria at two levels: title and abstract, followed by full-text. In total, 3,756 records were excluded at title and abstract level screening.

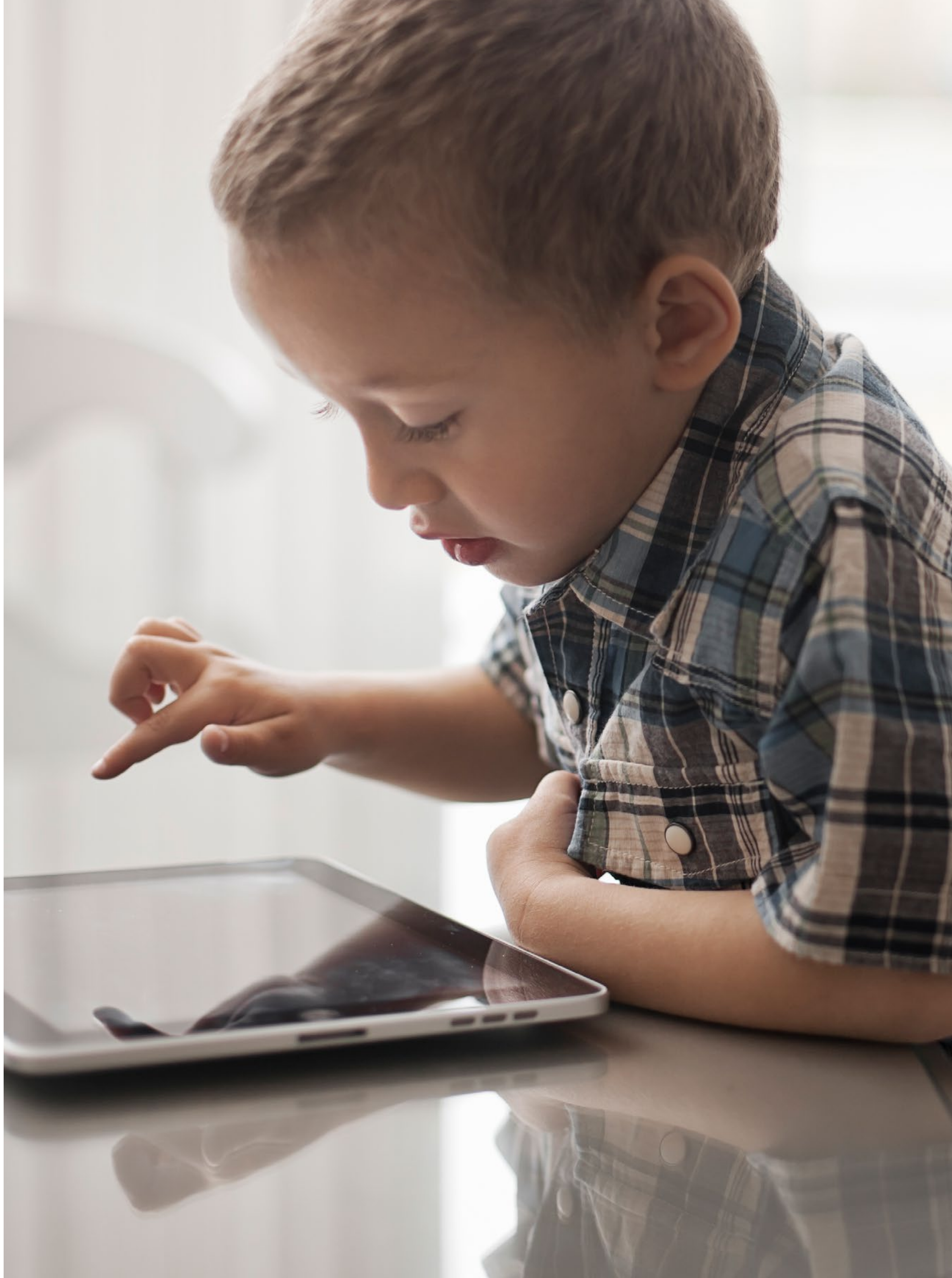
The remaining 126 studies were eligible for full-text screening. An additional two studies were included within the full-text screening, following contact with an author (n = 1) and the backward citation search of existing systematic reviews (n = 1). In line with the pre-registered protocol, full-text screening was also supplemented through contact with authors when more information to determine study eligibility was necessary (n = 22). Following this procedure, 83 records were excluded at full-text screening. Reasons for exclusion are provided in Figure 1. In total, 45 studies were identified for inclusion in the current review.

The prospective forward citation search of these 45 included studies returned 1,316 records. 1,249 records were excluded at title and abstract screening. The remaining 67 studies were assessed at full-text screening. At this stage, an additional five studies were identified for inclusion, resulting in a total of 50 studies included in the thematic narrative synthesis.

One reviewer (EE) was responsible for screening all records at both levels, with a second reviewer (LO) validating a random 20% sample (at both levels) to reduce potential error and bias in the screening outcomes. When conflicting decisions were apparent, discussions were held between the two reviewers until a consensus was reached. Inter-rater reliability between the two reviewers was calculated and reflected excellent agreement (κ = .83).

Figure 1: PRISMA Flow Diagram of studies through the systematic review





Quality of studies

Studies were not excluded from the synthesis based on quality to gain important insights into the impact of varied methodologies, as well as obtain an indication of the rigour of current research.

The quality of each included study was formally assessed using the study quality checklists developed by Kmet et al. (2004). Quantitative studies were scored (0 for no, 1 for partial, and 2 for yes) on 14 criteria, such as sampling strategy, appropriateness of study design, and group allocation procedures. Individual studies were then given an overall quality score (total score divided by the highest possible score of 28). Following Kmet et al.'s (2004) guidelines, studies were rated as low (0 – 0.44), moderate (0.45 – 0.69), and high (0.70 – 1.00) quality.

The same procedure was implemented for qualitative studies using 10 criteria, including systematic data analysis, reflexivity, and connection to a theoretical framework. If studies included both quantitative and qualitative components, both scoring checklists were used, and an overall quality score was derived from the mean average of the two (quantitative and qualitative) scores. Study quality scores did not affect inclusion in this review; all studies that met the inclusion criteria were subject to data extraction and thematic narrative synthesis.

Attrition rates and the reliable and consistent reporting of data within included studies were also used as indicators of study quality.

Results and Discussion

What type, frequency, and quality of evaluation studies have been conducted with educational maths apps?

Fifty studies were identified that evaluated 77 maths apps and included both quantitative and qualitative findings. A short descriptive account of each study is included in Appendix 1. A summary of the 50 included studies is also included in Appendix 2.

Experimental methods were the most common designs, with 20 RCTs and 13 quasi-experimental designs (QEDs). Four of the RCTs included an implementation process evaluation, of which one was reported in a separate paper. Three studies used other quantitative methods, including single case designs. There were seven qualitative studies and six using mixed methods.

Twenty-seven of the included studies were classified as high quality, based on Kmet et al.'s (2004) criteria. The remaining 23 studies were classified as moderate quality. No studies were classified as low quality. Only seven studies reported attrition rates greater than 30%.

However, the quality of reporting standards across the 50 studies was relatively poor. Fifteen studies did not sufficiently report data to enable calculations of within-subject effect sizes (Cohen's *d*) on children's learning gains in response to the maths app interventions. Similarly, time on task (i.e., the amount of time that children used the maths app interventions) was not reliably or consistently reported across studies.

What lessons can be learnt from previous research examining educational maths apps, relating to methods, populations, outcomes, and questions asked?

Methods

Business-as-usual control groups

Of the 12 experimental studies with a business-as-usual control group, eight reported positive results in favour of the maths app interventions compared to standard mathematical practice (Kosko & Ferdig, 2016; Nunes et al., 2019; Outhwaite et al., 2017; Pitchford et al., 2019; Schacter & Jo, 2016; Spencer, 2013; Stubbé et al., 2016; Wu, 2020). Standard mathematical practice in these studies incorporated mathematical instruction typical for the classroom context and included, whole class teacher-led instruction, small group or one-to-one activities, play-based learning, and traditional, physical manipulatives. However, the specific mathematical activities completed by the business-as-usual control group were not explicitly differentiated.

The remaining four studies comparing the maths app intervention to standard practice reported mixed findings. Specifically, the training effects of the maths apps were limited to the targeted skills, such as number line estimation, magnitude comparison, and spatial skills, and did not transfer to broader mathematical competence (Cornu et al., 2019; Lee & Choi, 2020; Zhang et al., 2020) or maths anxiety (Vanbecelaere et al., 2020).

Active control groups

Thirteen studies used a range of active control groups in an experimental design (RCTs and QEDs). Five studies compared the maths app interventions to non-digital maths interventions (Grimes et al., 2020; Mattoon et al., 2015; Miller, 2018; Schacter & Jo, 2017; Zander et al., 2016). Importantly, unlike in the business-as-usual control groups, the different mathematical activities, and thus the potential mechanisms for learning, were differentiated. There were mixed outcomes across these five studies.

When comparing the maths app intervention to one-to-one and small peer group instruction, results showed immediate, near-transfer benefits for mathematical performance (Grimes et al., 2020; Schacter & Jo, 2017). However, the maths app intervention effects did not transfer to mathematical language skills (Grimes et al., 2020). The benefits of app-based maths instruction were also seen when children used a paper-based version of the task, and then the app-based version. Training effects were not observed when the app was used first, followed by the paper-based task (Zander et al., 2016). Furthermore, learning gains with the maths app interventions were not statistically significant from gains made with traditional manipulatives and play-based learning (Mattoon et al., 2015; Miller, 2018). However, in both studies multiple apps were evaluated and the results per app were not disaggregated. As such, it is not clear which apps may or may not have supported learning. In addition, the frequency and duration of use for each app were not reported, thus it is also unclear which apps children engaged with and for long, which may have also impacted the observed results.

Eight studies compared the maths app intervention to other apps including maths (Hung et al., 2015; Judd & Klingberg, 2021; Schacter et al., 2016), reading/literacy (Berkowitz et al., 2015; Schaeffer et al., 2018), science (Parks & Tortorelli, 2020), and no educational or mathematical content (i.e., placebo) (Griffith et al., 2019; Hieftje et al., 2017). As the touch-screen tablet mode of delivery is consistent across the intervention and control groups in these studies, it enables the quality of the mathematical content to be assessed, rather than just the use of technology (i.e., a novelty effect).

Of these eight studies, six reported positive results in favour of the maths app interventions compared to the active controls (Berkowitz et al., 2015; Griffith et al., 2019; Hung et al., 2015; Judd & Klingberg, 2021; Schacter et al., 2016; Schaeffer et al., 2018). One study reported mixed results (Hieftje et al., 2017). Although the maths app intervention was shown to support children's numeration skills targeted by the app, the training effect did not transfer to other areas of mathematical development, including measurement, numerical operations, and problem solving (Hieftje et al., 2017).

In contrast, Parks and Tortorelli (2020) found no difference in mathematical learning gains between the intervention (nine maths apps) and active control group (five other educational apps covering literacy, science, and maths). However, in this study the effects of each individual app were not disaggregated, and both the intervention and active control groups received some form of app-based mathematics instruction.

Multiple control groups

Eight experimental studies (RCTs and QEDs) included more than one control group. Two studies included a business-as-usual control group, as well as an active control group, consisting of other apps with reading (Hassler Hallstedt et al., 2018) or no educational content (Pitchford, 2015). This design enabled the effects of the maths app intervention, as a form of effective instruction, to be disentangled from maturation (business-as-usual) and the technology-based delivery (active controls) within the context of the same study. Significant, near-transfer learning gains in response to the maths app interventions were observed in both studies. While Pitchford (2015) found these benefits transferred to mathematical conceptual knowledge, not targeted by the intervention, Hassler Hallstedt et al. (2018) found no significant improvements in far-transfer outcomes between the different groups.

Ramani et al. (2019) also compared the maths app intervention to two active control groups, consisting of an app-based working memory and reading game. As domain-specific mathematical skills and domain-general working memory skills are closely intertwined (Allen et al., 2019), this design allowed closer examination of near- and far-transfer effects of app-based instruction. Results showed children using the maths apps improved in numerical knowledge (near-transfer) compared to both active controls, as well as working memory (far-transfer), compared to the active controls using the reading game.

Five experimental studies used multiple control groups to compare different forms of implementation of the maths app intervention to understand how learning gains could be maximised. This included variations in time spent learning maths (Cary et al., 2020; Outhwaite et al., 2018), the language of instruction (Outhwaite et al., 2020), integration with other mathematical instructional materials (Pires et al., 2019), and additional links with parents at home (Schenke et al., 2020).

Results showed children's learning gains with maths apps were optimised when children used the apps for a longer, rather than shorter, duration (Cary et al., 2020). However, finding an appropriate balance with other subject areas is important to consider. Other research showed an extra 30 minutes per day learning mathematics, in the form of the app-based instruction had no additional benefit for children's basic and higher-order maths skills, compared to when the maths app intervention was integrated into the school day and thus did not take away from time spent with other subjects (Outhwaite et al., 2018).

Children also benefited more from the maths app intervention, when it was implemented in their first language (L1), compared to their second language (L2; Outhwaite et al., 2020), as well as when it was used in combination with physical play manipulatives, rather than virtual play manipulatives (Pires et al., 2019). However, there was no added benefit of implementing the maths app intervention with children alongside a parent-focused companion app, compared to the maths app only (Schenke et al., 2020).

Implementation process evaluations

Implementation process evaluations (IPEs) are typically used in combination with experimental methods (RCTs and QEDs) to further examine how a particular intervention works and under what circumstances (Humphrey et al., 2016; Pawson & Tilley, 1997). Within the 50 studies identified in this review, four RCTs included IPEs, which ranged from descriptions of implementation (Miller, 2018) to highlighting important implications for understanding how maths apps can be most effectively implemented to maximise children's learning outcomes (Nunes et al., 2019; Outhwaite et al., 2019) and understanding null results (Parks & Tortorelli, 2020).

Outhwaite et al. (2019) conducted observations of the maths app intervention sessions and interviews with the participating teachers and found 41% of the variance in children's learning outcomes (reported in Outhwaite et al., 2018) were explained by the established routine within the participating school contexts. This included implementing the intervention at a consistent time each day, having a member of staff responsible for the intervention implementation, having well-organised equipment, and having dedicated space within the classroom and a seating plan for the maths app intervention. In the larger-scale efficacy school-level RCT of the same maths app intervention, Nunes et al. (2019) triangulated qualitative data from observations, interviews, and questionnaires with teaching assistants. Results showed maths learning outcomes were greater when the teaching assistants perceived their role as an educator or guide, compared to an observer during implementation.

Similarly, when explaining their null findings, Parks and Tortorelli (2020) suggested that factors, such as lack of integration into the classroom setting, combined with a relatively short time on task (on average, 13 minutes per week) may have limited the success of the maths app intervention. Teachers also reported challenges with the logistics of charging tablet devices, difficulty downloading apps, and the need for training on how to integrate and implement the maths apps.

Feasibility studies

In the current review, there were seven feasibility studies with typically developing children with mixed results. These studies are valuable for establishing if the maths app interventions were viable (Kalmpourtzis, 2014; Stacy et al., 2017) or not (Swicegood, 2015), and provided the initial evidence base needed for further larger-scale experimental studies (Outhwaite et al., 2017).

These studies also provide useful insights into age-related differences in children's engagement behaviours with the maths apps. Moyer-Packenham et al. (2016) and Watts et al. (2016) found 5-8-year-old children's efficiency and accurate performance in different mathematical skills significantly improved over time, in response to using a selection of 11 maths apps. In contrast, Bullock et al. (2017) assessed the feasibility of three of these apps with 4-year-old children and found most children did not make any significant progress in seriation and counting skills.

Populations

Most studies focused on the use of educational maths apps with typically developing children of all attainment levels (n = 43). A small number of studies considered whether educational maths apps can support children underachieving in mathematics (n = 8) and children with special educational needs and disabilities (n = 3).

Children identified as underachieving in mathematics

Eight studies included a focus on children identified as underachieving in mathematics (Cary et al., 2020; Hassler Hallstedt et al., 2018; Hieftje et al., 2017; Kromminga & Codding, 2020; Nunes et al., 2019; Outhwaite et al., 2017; Outhwaite et al., 2018; Wu, 2020). The methods used to identify these children varied across studies, which has important implications for the internal validity of the findings.

Four studies included exploratory sub-sample analyses to examine which children benefited the most from the maths app intervention (Cary et al., 2020; Hieftje et al., 2017; Outhwaite et al., 2018; Wu, 2020). In these studies, children were identified as underachieving in mathematics based on a statistical cut-off point applied to their pre-test maths assessment score. For example, Cary et al. (2020) analysed a sub-sample of participants when evaluating *KinderTEK*. 5-6-year-old children performing below age-expected levels in mathematics (n = 38) made significantly stronger learning gains with the maths app intervention (Cohen's *d* = 1.97), compared to the control group. In contrast, children considered at-risk of being below age-expected levels in mathematics (n = 50) showed similar levels of improvement across the two experimental conditions. Similarly, both Outhwaite et al. (2018) and Wu (2020) found children identified as underachieving in mathematics made significantly greater improvements with *onebillion Maths 3-5 and Maths 4-6* (Cohen's *d* = 4.03), and *MathemAntics* (Cohen's *d* = 2.11), respectively, compared to their higher-attaining peers. Hieftje et al. (2017) also found the same pattern of results when examining *Knowledge Battle*. Collectively, this evidence may suggest that the maths app interventions were most beneficial for children underachieving in mathematics. However, this method of identifying children who are underachieving is not considered rigorous, due to the regression to the mean phenomenon and can threaten the internal validity of the findings (Barnett et al., 2004).

With a more rigorous approach, four studies used different measures to identify children as underachieving in mathematics, relative to their peers and evaluate the primary outcome of the maths app intervention. In the current review, two studies used teacher reports (Outhwaite et al., 2017; Nunes et al., 2019). One study used a screening assessment tool (Hassler Hallstedt et al., 2018) but this tool also included items used in the primary outcome measure that may have also been impacted by the regression to the mean phenomenon. One study used a combination of teacher reports and an independent screening assessment tool (Kromminga & Codding, 2020). Three studies found children using the maths app intervention made significant gains in maths skills compared to the control conditions (Hassler Hallstedt et al., 2018; Outhwaite et al., 2017; Nunes et al., 2019). Kromminga and Codding (2020) found children underachieving in mathematics made significant gains in maths skills with both the app- and paper-based versions of intervention.

Special educational needs and disabilities

Three studies assessed the feasibility of maths app interventions with children with special educational needs and disabilities (SEND), including Down syndrome and attention deficit hyperactivity disorder (ADHD), as well as vision loss, and emotional, behavioural, communication and learning difficulties (Ahmad et al., 2014; Pecora, 2015; Pitchford et al., 2018).

Within the included studies, most children with SEND showed improvements in mathematical skills in response to app-based instruction. However, the average progress rate was twice as slow for children with SEND relative to their mainstream peers. Moreover, the extent of children's additional needs significantly predicted their progress rate; children with hearing and/or language difficulties made slower progress compared to other SEND profiles (Pitchford et al., 2018). Further observations also indicated that children with SEND sometimes faced challenges engaging with the maths apps, such as becoming frequently distracted, lack of interest, unfocused, and interrupted by schedule changes (Ahmad et al., 2014; Pecora, 2015).

Outcomes

Mathematical outcomes

Included studies mostly focused on mathematical learning outcomes (n = 46), which were primarily assessed with researcher developed outcome measures (n = 32). Only 14 studies used a standardised assessment of mathematical attainment as the primary outcome measure (see Appendix 2).

Twenty-four studies sufficiently reported data to calculate within-subject effect sizes (Cohen's *d*) on children's learning gains in response to the maths app interventions and afford comparisons on the magnitude of effect sizes across studies. Effect sizes ranged from .05 to 3.34, with 11 studies reporting Cohen's *d* effect sizes equal to or greater than one (Grimes et al., 2020; Hassler Hallstedt et al., 2018; Kosko & Ferdig, 2016; Mattoon et al., 2015; Outhwaite et al., 2017; Outhwaite et al., 2020; Pitchford, 2015; Schacter & Jo, 2016; Vanbecelaere et al., 2020; Wu, 2020; Zhang et al., 2020).

Although most of these 11 studies used an RCT design (n = 6; Grimes et al., 2020; Hassler Hallstedt et al., 2018; Kosko & Ferdig, 2016; Pitchford, 2015; Vanbecelaere et al., 2020; Wu, 2020), studies with small sample sizes, such as those with fewer than 250 students, are more likely to produce inflated effect sizes than larger sample sizes (Cheung & Slavin, 2013). Of these 11 studies, nine had final sample sizes less than 250 children (Grimes et al., 2020; Kosko & Ferdig, 2016; Mattoon et al., 2015; Outhwaite et al., 2017; Outhwaite et al., 2020; Schacter & Jo, 2016; Vanbecelaere et al., 2020; Wu, 2020; Zhang et al., 2020). Hedge's *g* corrections were applied for studies with sample sizes equal to or less than 50 (Lin, 2018; see Appendix 2), and these effect sizes remained above one (Grimes et al., 2020; Kosko & Ferdig, 2016; Mattoon et al., 2015; Outhwaite et al., 2017).

Effect sizes can also be impacted by the outcome measure used. Of these 11 studies, seven used researcher developed measures as the primary outcome measure (Kosko & Ferdig, 2016; Outhwaite et al., 2017; Pitchford, 2015; Schacter & Jo, 2016; Vanbecelaere et al., 2020; Wu, 2020; Zhang et al., 2020), which are more likely to have larger effect sizes, compared to standardised assessment tools (Cheung & Slavin, 2013).

Only three studies used an RCT design with a large sample size (greater than 250 children) and a standardised measure of mathematical attainment (Berkowitz et al., 2015; Hassler Hallstedt et al., 2018; Outhwaite et al., 2018). Hassler Hallstedt et al. (2018) evaluated the *Chasing Planets* app for 19-20 weeks with 281 low-achieving students and reported an effect size of 1.19. However, there are important caveats to the internal validity of these findings due the crossover of items between the outcome measure and the measure used to identify children as underachieving in mathematics. As such, the large effect size may be inflated.

The two other studies with rigorous experimental designs reported effect sizes less than one. Berkowitz et al. (2015) evaluated the *Bedtime Math* app for 22 weeks with 278 children and reported an effect size of .82. Outhwaite et al. (2018) evaluated *onebillion Maths 3-5 and Maths 4-6* for 12 weeks with 389 children and reported an effect size of .78.

Long-term mathematical outcomes

Most included studies focused on mathematical outcomes measured immediately after the intervention period (n = 42). In the current review, five studies included a delayed post-test assessment of children's mathematical abilities with mixed results. Three studies showed the mathematical gains on the primary outcome measure were sustained at follow-up (Hassler Hallstedt et al., 2018; Outhwaite et al. 2017; Schaeffer et al., 2018). The timing of delayed post-tests in these studies ranged from five months (Outhwaite et al., 2017) to two years (Schaeffer et al., 2018). Two studies found the effects of the maths app intervention faded after one to (Ramani et al., 2019) two months (Vanbecelaere et al., 2020).

Time on task

Thirty-nine studies reported approximate time on task for the maths app interventions (see Appendix 2). For studies that reported time on task as a range (e.g., 600-1,350 minutes in Parks & Tortorelli, 2020), the intervention usage was estimated based on the mean average of the two reported values. On average, children used the maths apps for 797 minutes (13 hours), which ranged from 24 minutes (Ginsburg et al., 2019) to 5,400 minutes (90 hours; Stubbé et al., 2016). Within these 39 studies, 23 studies provided sufficient data to calculate within-group effect sizes. A small, positive, but not statistically significant relationship was observed between intervention usage and within-subject effect sizes on mathematical primary outcomes ($r = .30$, $p = .160$).

Use of in-app data

Four studies used log data automatically collected by the maths apps. For example, Hasanah et al. (2017) examined children's in-app behaviour, including the average number of steps made by children and the number of errors made. Pitchford et al. (2018) collected in-app data on progress through the app content and examined how this related to learning gains. Two studies used in-app data collection to measure children's mathematical abilities (Broda et al., 2019; Judd & Klingberg, 2021). This innovative use of data collection demonstrates proof-of-concept for conducting intervention studies remotely and for improving the quality of collecting data, particularly relating to usage and children's engagements with the apps.

Non-attainment outcomes

Alongside mathematical learning outcomes, 13 studies also included a range of non-attainment outcome measures, including enjoyment, preference modality, intrinsic motivation, and flow experience. Five studies reported that children mostly enjoyed using the evaluated maths app interventions (Berggren & Hedler, 2014; Ginsburg et al., 2019; Hieftje et al., 2017; Nunes et al., 2019; Pitchford et al., 2018). Furthermore, Hieftje et al. (2017) observed a positive relationship between enjoyment and children's learning gains in mathematics.

While Griffith et al. (2019) also found children enjoyed using the selection of seven maths apps, enjoyment ratings were not significantly different from that of other apps with no educational content. No significant differences were also observed in children's preferences between app-based learning and non-digital based maths interventions (Kromminga & Codding, 2020; Pecora, 2015; Swicegood, 2015).

Four studies examined the impact of educational maths apps on children's intrinsic motivation and flow experience (i.e., sustained attention) with mixed results. Hung et al. (2015) found greater flow experience, but not intrinsic motivation, with the maths app in the treatment group, compared to controls. The remaining three studies found no groups differences in intrinsic motivation or flow experience (Grimes et al., 2020; Spencer, 2013; Zander et al., 2016).

What is the external validity of the current evidence?

The external validity of the current evidence base was mixed. Although included studies were conducted in 18 countries, the majority were conducted in the USA (n = 26). Only eight studies examined the same maths app intervention. Between these eight studies three different countries were covered, including England (Nunes et al., 2019; Outhwaite et al., 2017; Outhwaite et al., 2018; Outhwaite et al., 2019), Brazil (Outhwaite et al., 2020), and Malawi (Pitchford, 2015; Pitchford et al., 2018; Pitchford et al., 2019).

However, in most included studies the maths app intervention was implemented in the classroom learning environment (n = 46) by teaching practitioners (n = 26), rather than by researchers (n = 20) (see Appendix 2). Only four studies were implemented in the home learning environment by parents (Berkowitz et al., 2015; Griffith et al., 2019; Kosko & Ferdig, 2016; Schaeffer et al., 2018).

What gaps can be identified based on the current evidence?

Across the 50 studies, eight key findings and associated gaps were identified in the evidence base. The contributions of studies to the eight themes identified through the thematic narrative synthesis is outlined in Appendix 3. These gaps should be addressed in future research to advance understandings of app-based mathematics instruction and ensure optimal learning outcomes for all children.

Near- and far-transfer benefits of educational maths apps

Nine experimental studies (RCTs or QEDs) explicitly examined differences between learning gains in mathematical skills targeted by the maths app intervention (near-transfer), and other relevant mathematical and cognitive outcomes, not included in the intervention (far-transfer) with mixed results. Three studies found that the benefits of the maths app interventions transferred to other skills (far-transfer), including higher-order and conceptual mathematical abilities, such as mathematical reasoning and problem solving (Outhwaite et al., 2018; Pitchford, 2015), as well as working memory (Ramani et al., 2019).

In contrast, six studies found the benefits of the maths app interventions did not transfer to broader mathematical abilities, including more complex arithmetic, problem-solving, and measurement skills (Cornu et al., 2019; Hassler Hallstedt et al., 2018; Hieftje et al., 2017; Lee & Choi, 2020; Vanbecelaere et al., 2020). Training effects also did not transfer to children's mathematical language skills (Grimes et al., 2020) or maths anxiety (Vanbecelaere et al., 2020).

As these studies represent the best available experimental evidence currently available, the mixed findings may indicate that variations in the breadth of the mathematical content and quality of the design features included in the evaluated apps may, in part, explain the variations in the observed near- and far-transfer training effects. However, no firm conclusions can be drawn based on the evidence to date.

Further research is needed to systematically evaluate these underlying pedagogical features within the maths apps to identify potential mechanisms that underpin children's learning in this technology-based context.

Children identified as underachieving in mathematics

While there is emerging evidence that educational maths apps can provide supplementary learning opportunities for children identified as underachieving in mathematics, the quality and rigour of this evidence was varied.

Five studies used the same assessment tool to identify children underachieving in mathematics, as well as measure the outcomes of the intervention (Cary et al., 2020; Hassler Hallstedt et al., 2018; Hieftje et al., 2017; Outhwaite et al., 2018; Wu, 2020). This poses a threat to the internal validity of these findings, due to the possibility of regression to the mean (Barnett et al., 2004). This is because extreme values, such as a low mathematics score at pre-test may be due to measurement error and are more likely to be followed by a less extreme score closer to the participant's true mean when assessed again at post-test (Barnett et al., 2004; Stigler, 1997). This makes it very difficult to disentangle a regression to the mean effect from a genuine intervention effect (Yudkin & Stratton, 1996).

In contrast, only three studies in the current review used an external method for identifying children as underachieving in mathematics, which was completely independent from the outcome variable (Kromminga & Codding, 2020; Outhwaite et al., 2017; Nunes et al., 2019). This helps to minimise the regression to the mean problem (Yudkin & Stratton, 1996), and thus greater confidence can be placed in these three studies, compared to the five using other identification methods. However, within these three studies, only one used a RCT design with a business-as-usual control group and a large sample size (Nunes et al., 2019), meaning that there is currently a lack of evidence as to whether educational maths apps are suited for children struggling to learn mathematics. As such, more large-scale rigorous intervention studies are needed to evaluate the use of educational maths apps with young children identified as underachieving in mathematics and inform the targeted implementation of such interventions. The use of an assessment measure for identifying these children, independent of the outcome measure, should be considered best practice when conducting future research.

Special educational needs and disabilities

Research with children with SEND was restricted to feasibility level evidence and no studies worked with children with dyscalculia. However, it is important to acknowledge that the inclusion criteria for the current review was restricted to the first three years of formal education. As many special educational needs are not formally diagnosed until children are older and the SEND code of practice ranges from 0-25 years (DfE, 2020), there may be more developed evidence that fell beyond the age range included in the scope of this review.

Nevertheless, further research is needed to examine the use of educational maths apps with children with specific mathematical difficulties, such as dyscalculia, Down syndrome, and other learning difficulties. For example, Sella et al. (2021) showed children with Down syndrome made significant improvements in specific numerical skills and mental calculation, in response to using the computer-based game, *The Number Race*, for 10 weeks, compared to a reading active control group. Although this study fell beyond the scope of the current review, it provides a useful demonstration of how interventions initially evaluated with typically developing children (Sella et al., 2016), can be translated for different population groups. As such, similar research specific to educational maths apps on touch-screen tablets is needed to identify how such interventions can be successfully implemented and adapted to meet the needs of different groups of children and ensure equitable access to effective and evidence-based mathematical instruction.

Role of age and language in educational maths apps

Consistent with previous systematic reviews (Griffith et al., 2020; Herodotou, 2018), the current synthesis also suggests that younger children may face more barriers when accessing educational maths apps (e.g., Bullock et al., 2017). However, this chronological approach, does not necessarily capture how the suitability of educational maths apps may differ based on children's individual abilities. Two studies showed children's language abilities influenced their progress through the maths app intervention, in that children with better proficiencies in the language of instruction made more progress through the maths apps, compared to those with weaker language skills (Pitchford et al., 2018; Outhwaite et al., 2020). Overall, this evidence collectively suggests that as young children are still developing their language and vocabulary skills, it may be important to consider their strengths and weaknesses in these areas, when deciding if and which apps to use, rather than just their age.

Furthermore, different types of apps may be better suited to support children's developing language skills in the context of app-based maths instruction. For example, apps that are designed with a strong focus on social interaction (Hirsh-Pasek et al., 2015), may provide additional linguistic scaffolds that are not included in other types of apps that place an emphasis on individual use by the child. As such, future research should further examine the role of language and how this may interact with age, when children use different types and features of maths apps.

Usage and immediate and sustained learning gains

Although a small, positive, but not statistically significant relationship was observed between intervention usage and learning outcomes ($r = .30$), this result should be treated tentatively. This is because usage was not reliably and consistently reported across studies and very few studies examined the log data available from the maths apps. In addition, the quality of reporting standards across the 50 studies was relatively poor, which limited the calculations of within-subject effect sizes on learning outcomes.

In studies where usage was sometimes adequately reported, it gave insights into explaining the observed long-term mathematical outcomes, in response to the maths app interventions. Three studies reported learning gains that were maintained when assessed at a 5-month- to- 2-year- follow-up (Hassler Hallstedt et al., 2018; Outhwaite et al. 2017; Schaeffer et al., 2018). Importantly, in these studies, the maths app intervention was implemented over a sustained period, with time on task ranging from 900 to 1,698 minutes (15- 28 hours). In contrast, the two studies that found the effects of the maths app intervention faded at one to two months later, were implemented for a shorter duration of 100-300 minutes (1.7- 5 hours) (Ramani et al., 2019; Vanbecelaere et al., 2020). Overall, this highlights the need for more reliable and consistent reporting of maths app usage, which can be supported by innovative methods for data collection.

Innovative methods for data collection

Previous systematic reviews on educational apps identified the lack of studies utilising in-app data on learner analytics (Herodotou, 2018). The current review identified four studies that used log data to understand children's in-app behaviour and examine learning outcomes (Broda et al., 2019; Hasanah et al., 2017; Judd & Klingberg, 2021; Pitchford et al., 2018). These studies demonstrate proof-of-concept for innovative methods for data collection with educational maths apps. Future studies should build on this evidence base to make effective use of the in-app data automatically generated by the maths apps. In particular, in-app data can be used to rigorously examine the relationship between time on task and mathematical outcomes, as well as understanding other areas of in-app behaviour, such as patterns of errors children make during game play. This data could also be utilised to conduct high-quality intervention studies remotely, for example during periods of home learning and for working with hard-to-reach populations, such as those with SEND.

Cross-cultural comparisons with educational maths apps

Although 18 different countries were represented across the 50 included studies in the current review, only one maths app intervention was evaluated in different high- (England), middle- (Brazil), and low- (Malawi) income country contexts (Nunes et al., 2019; Outhwaite et al., 2017; Outhwaite et al., 2018; Outhwaite et al., 2019; Outhwaite et al., 2020; Pitchford, 2015; Pitchford et al., 2018; Pitchford et al., 2019). Across these eight studies, results were positive in favour of the maths app intervention, compared to a range of control groups. As such, it demonstrates that this maths app intervention can be effectively used to support children's mathematical development, particularly for girls in countries where gender differences in standard practice may typically hinder their learning progress (Pitchford et al., 2019). This has important implications for addressing global educational challenges, including issues faced because of schools around the world being closed for significant periods of time for most children during the Covid-19 pandemic. Data from the United Nations (2021) suggests that the impact of the pandemic has wiped out 20 years of educational gains, thus limiting progress towards the education Sustainable Development Goals. To continue supporting children's learning and development, particularly in response to Covid-19, there has been a greater emphasis on the implementation of children's education by parents at home, for which technology can play an important role. However, it is essential these interventions are effective, evidence based, and distributed in an equitable way. In response, further rigorous research is needed to evaluate educational maths apps around the world, particularly in the home learning environment.

Educational maths apps in the home learning environment

The home learning environment plays a vital role in child development (Toth et al., 2020), yet parents typically engage in maths activities at home once a week, compared to every day for reading (Organisation for Economic Co-operation and Development [OECD], 2020). Parents also often report their own maths anxieties, which can impact their children's mathematical outcomes (Maloney et al., 2015).

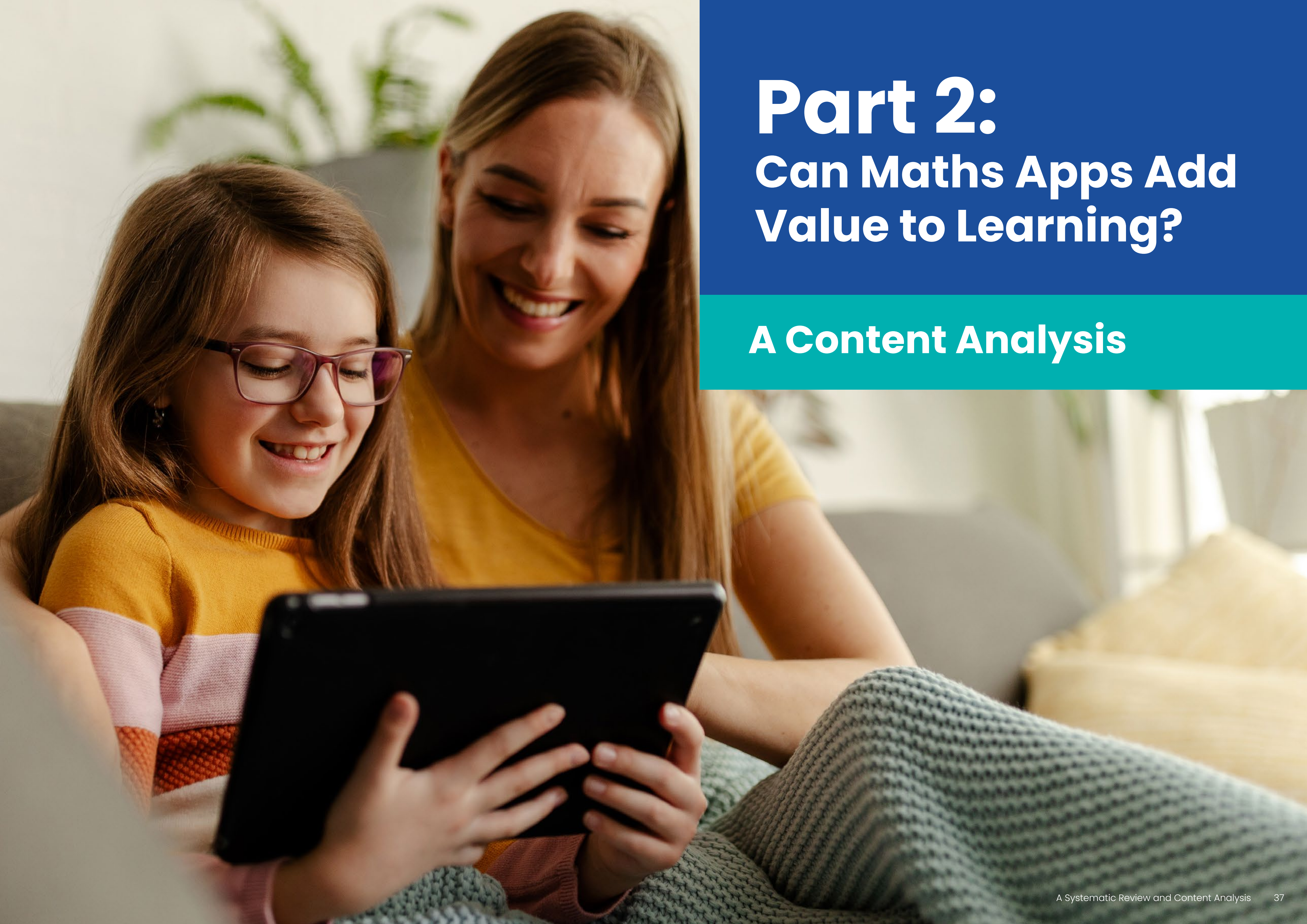
In the current review, only four studies were conducted at home with parents (Berkowitz et al., 2015; Griffith et al., 2019; Kosko & Ferdig, 2016; Schaeffer et al., 2018). One additional study involved parents in the maths app intervention implemented in the classroom (Schenke et al., 2020). As such, more research is needed to understand how maths app interventions can work in the home to empower parents and support children's learning. This could be achieved through apps that encourage off-screen play and maths talk (Berkowitz et al., 2015; Schaeffer et al., 2018). No apps in the current review included a separate parent area that communicated children's usage and progress through the app. This may support parental involvement in children's mathematical learning, but further research is needed.

Part 1 Conclusions

The current systematic review was inclusive of quantitative, qualitative, and mixed methods studies to comprehensively capture the emerging evidence on whether educational maths apps can support learning, under what circumstances apps work, and who benefits the most from apps. Most studies were implemented in the classroom by teaching practitioners with typically developing children, highlighting the usability and external validity of the evaluated maths apps for this group of children.

The best available experimental evidence (RCTs and QEDs) mostly demonstrated positive results towards the maths app intervention for typically developing children, compared to a range of control groups, including standard mathematical practice (business-as-usual), and other educational apps (active control).

However, only two of these studies met the highest standards for rigorous methods, including, a RCT design with a large sample size (greater than 250 children), a standardised measure of mathematical abilities, and sufficiently reported data to calculate within-subject effect sizes (Berkowitz et al., 2015; Outhwaite et al., 2018). As such, although the current evidence base demonstrates promising and externally valid findings, more high-quality rigorous research is needed to support evidence-based decisions relating to young children's use of educational maths apps.



Part 2:

Can Maths Apps Add Value to Learning?

A Content Analysis

Part 2: Can Maths Apps Add Value to Learning? A Content Analysis

This content analysis follows the systematic review synthesising the research literature on educational maths apps, reported in Part 1. The systematic review identified 77 maths apps that have been previously evaluated across 50 studies. Overall, the studies predominately report greater learning outcomes for children using the evaluated maths apps, compared to a range of control conditions. While this systematic review addresses whether educational maths app can be an effective learning tool, little research has considered what the active ingredients (i.e., mechanisms), or combination of ingredients, of successful maths apps are and how they link to current theories of mathematical development and learning science. To examine how different maths apps work, the underpinning pedagogy, including the mathematical content and app design features, needs to be examined (Griffith et al., 2020) and linked to the observed learning outcomes (Outhwaite et al., 2019).

The Need to Develop a New Content Analysis Framework

To date, only one systematic review on educational apps (maths and literacy) has attempted to consider the potential relationships between the app design features and observed learning outcomes. Kim et al. (2021) scored 36 identified apps in their systematic review using five simple questions based on Hirsh-Pasek et al.'s (2015) theoretical framework that high quality educational apps should include active, engaged, meaningful, and socially interactive learning with a specific learning goal. Questions included, "do activities promote meaningful learning?", which were then rated on a 3-point Likert scale. In this example, scores ranged from: "low (app promotes rote memorisation disconnected from broader learning contexts), moderate (app sometimes connects to broader learning contexts but also sometimes promotes rote memorisation), to high (app promotes conceptual mastery that is consistently connected to a broader learning context)" (Kim et al., 2021, p.4 Supplementary Materials). Moderation analyses showed no relationship between the quality of app score and the learning gains in the reported meta-analysis. However, the questions used to assess the apps did not include specific features, such as feedback or levelling, which are proposed in other existing frameworks to be important components of app-based instruction (Callaghan & Reich, 2018; Herodotou, 2021; Kolak et al 2020; Meyer et al., 2021; Papadakis et al., 2017).

Feedback within educational maths apps can be defined as immediate responses from the app, based on the actions, input, and performance of the user (Tucker, 2015). Feedback can be understood according to its explanatory and motivational components. Explanatory feedback within the apps provides the user with an explanation of why their answer is correct/incorrect. On the other hand, motivational feedback provides general feedback to the user such as "You did it!" or "Great job!" but is not directly associated with the answers or performance of the user. Recent experimental evidence shows preschool-aged children who received explanatory verbal feedback, made significantly fewer errors during a novel practice-based mathematics sorting game compared to children who received motivational non-verbal feedback (e.g., cheering sound effects). Motivational verbal feedback (e.g., "Great job!") did not increase performance accuracy compared to other forms of feedback (Callaghan & Reich, 2021).

Levelling can be defined as tailoring learning content so that it accounts for and builds on children's prior knowledge and progression (Hsin & Wu, 2011; Magliaro et al., 2005). Levelling can be implemented in educational maths apps in three ways. First, participatory free form levelling refers to apps that provide a suggested but not enforced sequence of learning content (Kucirkova, 2018). Second, programmatic static levelling tailors the learning content to a child based on an initial attainment assessment or learning content that is pre-selected by an adult (Vandewaetere & Clarebout, 2014). Third, programmatic dynamic levelling adapts the presented learning content in response to a child's performance while using the app

(Vandewaetere & Clarebout, 2014). Experimental evidence demonstrated app-based maths learning tasks were completed quicker and with increased accuracy when learning tasks gradually increased in difficulty (programmatic levelling), compared to a non-strategic sequence of learning activities (participatory free form levelling) (Callaghan & Reich, 2021). Further research found no group differences between dynamic and static forms of programmatic levelling (Vanbecelaere et al., 2021).

Within previous research, it is also assumed that all app design features, such as levelling and feedback, are equally important and combine in equivalent ways. However, some app design features may be necessary, sufficient, or inconsequential. Therefore, more nuanced approaches and analyses are required to understand how maths apps may work to support learning.

Existing Frameworks for the Educational Value of Apps

Many existing frameworks assessing the educational value of apps have been developed based on developmental and learning science theory (e.g., Hirsh-Pasek et al., 2015) and thus adopt a top-down approach to coding educational maths apps (Callaghan & Reich, 2018; Kolak et al 2020; Meyer et al., 2021; Papadakis et al., 2017). For example, Meyer et al. (2021) designed a detailed coding scheme and scoring system also based on Hirsh-Pasek et al.'s (2015) 4-pillar theoretical framework. However, many of the items included in Meyer et al.'s (2021) framework were biased towards certain types of apps (Kay & Kwak, 2018). For example, off-screen social interaction is a key component of parent-based apps (e.g., *Bedtime Math*) but is not aligned to the design principles of other types of apps, such as practice-based and game-based apps, which are typically designed to be individually used by the child. As such, when using Meyer et al.'s (2021) scoring scheme, parent-based apps will inherently gain a higher score compared to other types of apps. This is problematic, as no intervention studies to date have directly compared the learning outcomes of these different types of apps with these features. Thus, it is currently unknown whether this distinction between face-to-face interaction and in-app character engagement is appropriate within the context of app-based instruction.

A similar issue is also present in Kolak et al's (2020) framework, with the inclusion of the storyline items, which are a key component of game-based apps and are less relevant to other types of apps. To develop unbiased conclusions about the potential mechanisms underpinning learning outcomes with app-based maths instruction, it is important to develop framework items that are applicable to the different types of maths apps.

In contrast, Herodotou (2021) took a bottom-up approach and identified, through a literature review, app features that may facilitate or hinder learning across different subject areas including: main figure, feedback, instructions, highlighting information, constraints, linking multiple representations, experimentation, as well as other features such as, progression, sounds, and language (e.g., Fallon, 2013; Moyer-Packenham

et al., 2016). The proposed framework was then expanded through the observational study of 17 children using one practice-based maths app (*Moose Math*), which emphasised the importance of feedback, experimentation, and learning theories in app design.

Herodotou's (2021) proposed framework was based on a narrative review of studies across different domains and was not focused solely on maths apps. In the current project, we took this approach further by (a) conducting a systematic review of the literature focusing specifically on maths apps with young children (see Part 1). This resulted in identifying several maths apps that were shown to support learning and will be combined with (b) a content analysis of the identified apps and their design features.

This content analysis combined a top-down (deductive) and bottom-up (inductive) approach, which enabled the current study to build on existing frameworks and ensure other app features, not currently included in existing frameworks, are also captured within this analysis. Furthermore, while existing frameworks (Callaghan & Reich, 2018; Herodotou, 2021; Kolak et al 2020; Meyer et al., 2021; Papadakis et al., 2017) provide a useful indication of specific potential mechanisms underpinning app-based learning, the associated scoring systems are not suitable for making meaningful connections to intervention study outcomes. The current study addressed this limitation using Qualitative Comparative Analysis (QCA).

Research Questions

This content analysis aimed to address the following research questions:

RQ1: When combining top-down and bottom-up approaches to the content analysis, what mathematical content and app design features are included within educational maths apps that have been previously evaluated in terms of their impact on children's learning outcomes?

RQ2: Which app design features are sufficient for supporting children's learning outcomes with maths apps?

Method

Inclusion criteria

To be eligible for inclusion in the content analysis, the maths apps identified in the systematic review had to meet the following criteria:

- Included apps must be the individual focus of an intervention study that has mathematical attainment as the primary outcome measure, measured before (pre-test) and after (post-test) the intervention period (e.g., the *onebillion* Maths app intervention evaluated in Pitchford et al., 2018).
- In cases where multiple apps were included in one study, the results (i.e., mathematical attainment scores) were reported separately for each app (e.g., Bullock et al., 2017 evaluates three apps and includes individual results for each app).

- If the intervention study included multiple maths apps and the reported results were not disaggregated for each app (Griffith et al., 2019; Mattoon et al., 2015; Miller, 2018; Parks & Tortorelli, 2018), the identified maths apps were excluded from the content analysis. This is because it was not possible to disentangle whether one app was driving the observed learning outcomes and there may be some apps that were not used during the intervention period, as intervention intensity specific to each app was not reported. There was also no guarantee that these apps, with combined results, had the same app design features, therefore the identification of active ingredient/s would be unclear.
- However, before excluding maths apps that were included in intervention studies without disaggregated results, the lead author of each study was contacted to check whether any app specific learning outcome data was available.

These criteria allowed for within-subject effect sizes in response to the specific maths app to be calculated, and thus enable the mathematical content and app design features to be meaningfully associated with the observed learning outcomes (Outhwaite et al., 2019).

Included apps must also:

- Be commercially available and accessible for download from the Apple and/or Google Play stores in the UK.
- If the app was not commercially available (e.g., it has been developed by a research team for the purposes of the study), the lead author was contacted to request access. If it is not possible to access the relevant app after two follow-up emails, the app was excluded from the content analysis.
- If the app is not available in the UK app stores, the app was excluded from the content analysis.

Effect sizes

To capture children's learning outcomes with the evaluated maths apps that met the inclusion criteria, the following data was extracted from the relevant studies:

- Group mean and standard deviation for pre-test and post-test mathematical attainment scores for the intervention group(s).
- Final sample size of the intervention group(s).
- Intervention intensity: number of weeks the intervention was implemented for and the number and length of sessions per week.
- Whether the mathematical assessment tool used as the outcome measure was standardised or researcher developed.

The extracted data on pre-test and post-test mathematical attainment scores was used to calculate within-subject (i.e., pre-test to post-test) effect sizes (Cohen's *d*) for the progress made in maths abilities over the duration of the intervention period for the intervention group only. The final sample size was used to apply Hedge's *g* corrections where appropriate for studies with samples less than 50 (Lin., 2018).

Coding of included apps

Each of the included apps were coded following a three-step process, as detailed below.

Type of maths app

First, each included app was classified by type. This was based on Kay and Kwak's (2018) taxonomy of different types of educational apps. While these categories were exclusive, they were not exhaustive; an important feature of qualitative coding (Braun & Clarke, 2006). As such, the taxonomy was expanded to include parent-based apps (see Table 3).

Table 3: Descriptions for each type of educational app.

Type of App	Description
Practice-based	Designed to support the acquisition of learning content, such as mathematical facts and concepts, through targeted practice (e.g., *onebillion Maths* 3-5). Primarily designed to be used by children individually and is self-paced. Child is the consumer of the learning content.
Game-based	Same as practice-based apps, with the addition that learning content is embedded within broader immersive player narrative (e.g., *Slice Fractions*).
Constructive	Designed to encourage the exploration and active manipulation of mathematical ideas and concepts (e.g., *Montessori Numbers for Kids*). Child is the consumer of the learning content.
Productive	Designed to support children to produce their own content, for example to present their own ideas on a particular maths topic; they are creators of their own learning content (e.g., *Quizlet Plus*).
Parent-based	App content is primarily designed for parents/caregivers and encourages offline interactions and learning opportunities with children (e.g., *Bedtime Math*).

Mathematical content included within the app

Second, the mathematical content within each included app was catalogued, based on four areas of mathematical development which included several description points: Number Representation and Relationships (11 description points including transcoding, number bonds, and number line estimation), Counting (7 description points including one-to-one correspondence, cardinality, and skip counting), Arithmetic (10 description points including addition, subtraction, and arithmetic symbols and language), and Shape, Patterns, and Measurement (10 description points including working with patterns, shape recognition, and sequence of events).

The initial categories were developed based on current theories of mathematical development (Butterworth, 2005; Clements & Sarama, 2009; Gilmore et al., 2018) and best practice guidelines for effective early mathematics teaching (Clark et al., 2020). The descriptions within each area of mathematical development also aligned with content from the mathematics curriculum in England (DfE, 2013; 2021) and the Common Core Standards for mathematics in the USA (Common Core State Standards Initiative, n.d.).

Included apps were dichotomously coded (not present [0] or present [1]) for each of the listed areas of mathematical development. Apps needed to meet at least one of the description points, for the area of mathematical development to be coded as present.

To capture more detail about how well the different areas of mathematical development were covered within the included apps, each of the description points were also dichotomously coded (not present [0] or present [1]). There was also the opportunity to include other mathematical skills, not currently listed, to ensure exhaustive coding.

Design features included within the app

Finally, the presence of certain app design features was dichotomously coded (not present [0] or present [1]). This was followed by a brief description of how each feature was implemented within each app. Table 4 outlines the list of app design features identified. This list was collated based on their predominant presence in existing frameworks evaluating the educational value of apps (Callaghan & Reich, 2018; Herodotou, 2021; Kolak et al 2020; Meyer et al., 2021; Papadakis et al., 2017).

However, as discussed above, most of the existing frameworks adopted a top-down approach, which is not exhaustive and there may be additional design features present within the included maths apps that are not currently included. As such, a bottom-up approach was also taken with the opportunity to include other design features – beyond those outlined in Table 4 - to ensure exhaustive coding. Additional design features were identified through reviewer engagement with the app and by re-visiting the associated papers and applying the principles of an intervention component analysis (Sutcliffe et al., 2015).

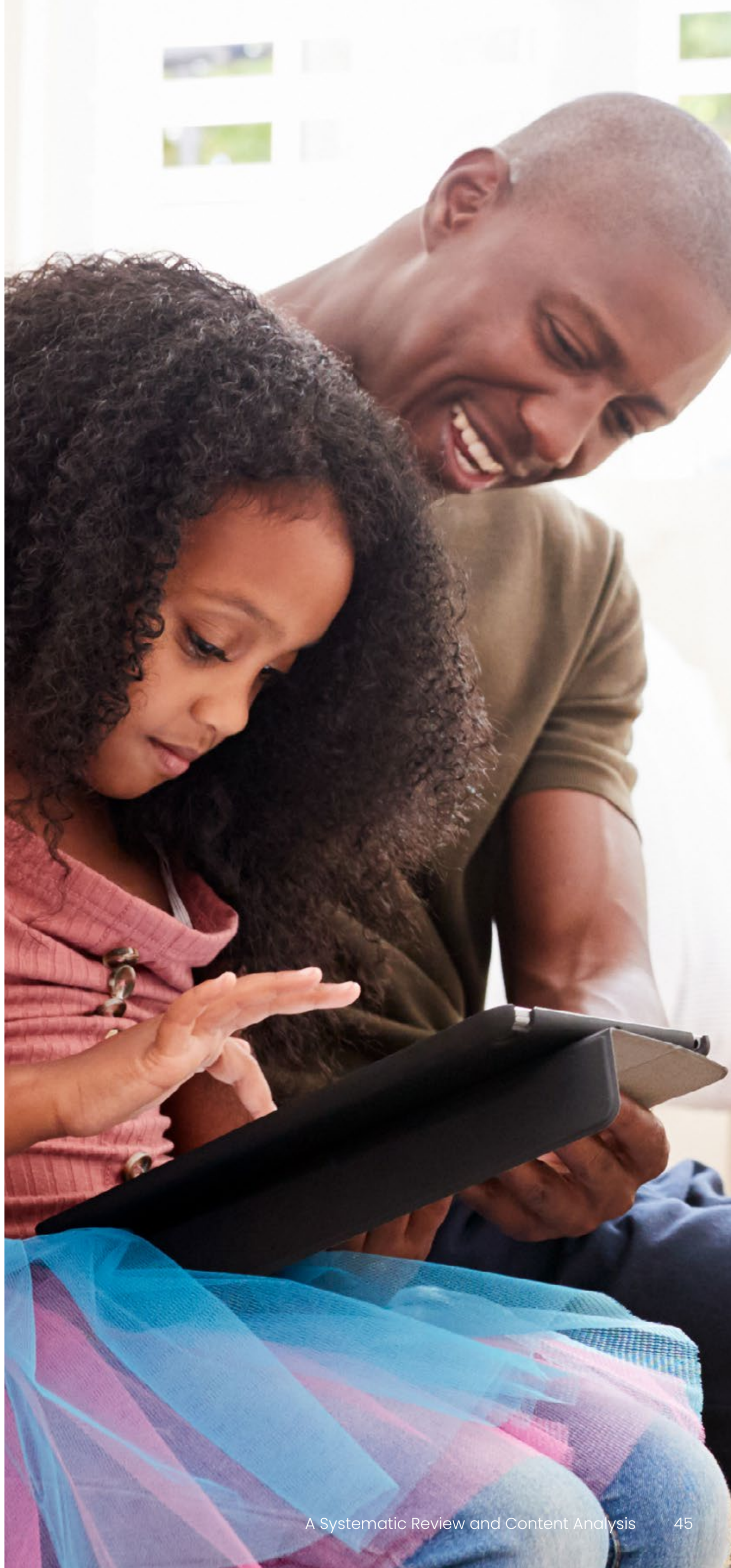


Table 4: Initial list of design features

Design Feature	Description	Inclusion in Existing Frameworks				
		1	2	3	4	5
Feedback	App provides immediate feedback on children's performance during the learning activity. Nature of feedback can be: 1) **Explanatory/corrective:** app gives reasons why an answer is (in)correct, and/or 2)**Motivational:** app encourages children's engagement (e.g., "Good job!" or "Let's try again"). Feedback can be delivered via audio and/or visually.	x	x	x	x	x
Levelling	App provides multiple, structured levels of difficulty that can be personalised to the child. Levelling can be: 1) **Programmatic, static:** learning content is tailored to the child based on initial attainment assessment or pre-selected by an adult, and/ or 2) **Programmatic, dynamic:** adaptivity of learning content in response to the child's performance, or 3) **Participatory, free form:** App provides suggested (but not enforced) sequence of learning content.	x	x	x	x	x
Social interaction	App encourages social interaction either: 1) With an in-app character, who 'communicates' with the child through introducing the learning activity, modelling concepts and/or asking prompting questions. 2) Between the adult and child outside of the app. Support is given to the adult to support these interactions.		x	x	x	x
Task instruction	App includes: 1) Explicit instructions (i.e., explains task and what to do next) with easy-to-understand language, and/or 2) There are opportunities for instructions to be repeated.		x	x	x	x
Meaningful learning and solving problems	Activities within the app provide meaningful learning centred around mathematical development and focus on: 1) Practicing basic mathematical skills in isolation (e.g., tracing Arabic digits only), and/or 2) Practicing multiple basic mathematical skills in relation with each other (e.g., tracing Arabic digits and transcoding with other numerical representations), and/or 3) Embedding mathematical skill practice within a real-life context (e.g., selecting and counting items to be packed on holiday), and/or 4) Applying mathematical skill practice to solve novel problems.		x	x	x	x

Existing frameworks: 1) Callaghan & Reich, 2018; 2) Herodotou, 2021; 3) Kolak et al., 2020; 4) Meyer et al., 2021; 5) Papadakis et al., 2017.

Coding procedures

Each of the included apps was coded using the data collection form designed specifically for this study (see Appendix 4). To assess the mathematical content, reviewers considered the overall focus and available learning tasks for each app. To understand the app design features, reviewers played with each app for 20-30 minutes and engaged with a minimum of ten different activities from across the app. During play, reviewers gave both correct and incorrect answers to understand how each maths app responded to user behaviour.

The first reviewer (EE) completed the coding procedure for all the included maths apps. A second reviewer (LO) repeated the coding procedure for 20% (randomly selected) of the apps. There was excellent agreement between the reviewers (κ= .85). Any disagreements were resolved through discussion.

Results

Type of apps

As outlined in Table 5, practice-based apps were the most popular type of app that has been evaluated. Productive and parent-based apps were the least common.

Table 5: Summary of the type of apps identified in the systematic review (Part 1; n = 23).

Type of App	Number of Apps	App Names
Practice-based	15	*Fingu; Friends of Ten; Intro to Math; IXL; KinderTEK; Know Number Lite; MathemAntics; Maths 3-5; Maths 4-6; Math Shelf; Montessori Numbers for Kids; Native Numbers; Pink Tower; Splash Math 2nd Grade; Teaching Number Lines*
Game-based	4	*Addimal Adventure; Slice Fractions; Vektor; Zorbit's Math Adventure*
Constructive	2	*Montessori Bead Skip Counting; 100s Board*
Productive	1	*Quizlet Plus*
Parent-based	1	*Bedtime Math*

Mathematical content

Results showed that within the 23 evaluated maths apps, understanding number representations and relationships was the most popular area of mathematical development (n = 21). There were 18 apps that targeted counting skills, 12 apps targeted arithmetic, and 13 apps targeted shape, patterns, and measurement. Table 6 summarises the specific mathematical skills covered within these areas of mathematical development.

Four apps also included additional mathematical areas that were not categorised in the initial coding (*IXL; Slice Fractions; Splash Maths 2nd Grade; Vektor*). *IXL* included the exploration of probability, data, and graphs. *Slice Fractions* included the addition and subtraction of fractions. *Splash Maths 2nd Grade* explored how to read data from line graphs, picture graphs, and bar graphs. *Vektor* included visuo-spatial working memory tasks and shape rotation tasks.

Table 6: Summary of the mathematical content of the included maths apps (n = 23)

Mathematical Content	Number of Apps
Number Representation and Relationships	
Number representation with Arabic digits (symbolic), verbal, and written number word recognition.	19
Transcoding between Arabic digits and number words.	10
Recognition of number of items within a set (i.e., subitising).	14
Comparison of magnitude of symbolic (i.e., 5 is less than 8) and non-symbolic number (i.e., 5 dots is less than 8 dots).	9
Mathematical language, relating to more/ less, before/ after, most/least, equal to.	10
Position of numbers on a number line (e.g., 0-10, 0-20, and/or 0-100).	13
Number bonds to 10 and/or 100.	5
Odd and even numbers.	5
Basic exploration of fractions (e.g., recognition of halves and quarters).	4
Further exploration of fractions (e.g., ½ of 6 = 3 and/or equivalence of ½ and ²⁄₄).	4
Exploration of place value (i.e., understanding the value of a digit based on its position in a number).	4
Counting	
Support for one-to-one correspondence.	18
Counting sequence (in ones) forwards and backwards to 10, 20 and/or 100.	12
Support for cardinality- the last number counted represents the set.	14
Counting of a range of different objects (i.e., understands oneness, twoness, threeness with abstraction principle).	11
Counting on from a specified number (e.g., count from 5 to 8).	10
Skip counting (e.g., in twos, threes, fives, and/or tens).	6
Given a number, identify one more and one less.	5

Mathematical Content	Number of Apps
Arithmetic	
Single digit addition and subtraction with two numbers (e.g., 2 + 2 = 4).	12
Addition and subtraction with at least one double-digit (e.g., 10 + 2 = 12 or 10 + 10 = 20).	10
Single digit addition with three numbers (e.g., 1 + 2 + 3 = 6).	3
Familiarisation with arithmetic (e.g., + - =) and/or mathematical symbols (e.g., < >) and associated language (e.g., add, take-away, equals, more than, less than).	10
Familiarisation of how arithmetic operations are related to each other (e.g., + is opposite of -).	7
Introduction to multiplication tables (e.g., 2, 5, and 10).	2
Support for counting strategies when completing more complex arithmetic operations (e.g., share and group objects).	10
One-step addition and subtraction word problems using concrete objects, pictorial representations, and/or missing number problems.	9
One-step multiplication (e.g., doubling) and division (e.g., halving) word problems using concrete objects, pictorial representations and or arrays.	3
Familiarisation of commutative nature of addition and/ or multiplication (i.e., numbers can be added/ multiplied in any order) and non-commutative nature of subtraction and division.	6
Shape, Patterns, and Measurement	
Recognise and continue patterns with shapes and/or objects (i.e., not number).	7
Create new patterns with shapes and/or objects.	4
Order objects in a set (i.e., first, second, third).	6
Sequence of events in chronological order.	1
Compare or group sets of objects on a defining feature (e.g., all the blue balls).	7
2D and/or 3D shape recognition.	6
Understanding of 2D and/or 3D shape properties (e.g., triangle has 3 corners, line of symmetry).	6
Manipulate shapes to make and/or decompose other shapes.	4
Mathematical language relating to length, height, weight, capacity, spatial position (e.g., left, right, top, bottom), money, and time.	6
Use of measurement units (e.g., m/cm, kg/g, litres/ml, £/p).	4

App design features

As shown in Table 7, within the 23 apps, explanatory and motivational feedback (n = 12) and participatory, free form levelling were the most common (n = 12). Most apps did not provide social interaction in the form of an in-app character or support adult-child interactions outside of the app.

All apps provided some form of meaningful learning and problem solving with most apps including practice of basic mathematical skills in isolation (n = 17). Of the two apps that incorporated a mixture of meaningful learning and problem-solving features, *Bedtime Math* practiced multiple basic mathematical skills in relation with each other and embedded mathematical skill practice within a real-life context, whilst *IXL* practiced basic mathematical skills in isolation and embedded mathematical skill practice within a real-life context. *Quizlet Plus* was categorised as not applicable as it is a productive app and as such children are creators of the app content.

Table 7: Summary of the app design features of the included maths apps (n = 23)

App Design Feature	Number of Apps
Feedback	
Explanatory and motivational feedback	12
Explanatory only	0
Motivational only	8
No feedback	3
Levelling	
Participatory, free form	12
Programmatic, dynamic	5
Programmatic, static	4
No levelling	2
Social Interaction	
In-app character	10
Adult-child interactions	2
No social interaction	11
Task Instruction	
Instructions can be repeated by the child more than once in all app activities	6
Instructions are repeated only once if a child does not respond or gives an incorrect answer	3
Instructions can be repeated, but only in some app activities	2
Instructions cannot be repeated	7
No instructions	5
Meaningful Learning and Solving Problems	
Practicing basic mathematical skills in isolation	17
Practicing multiple basic mathematical skills in relation with each other	3
Embedding mathematical skill practice within a real-life context	0
Applying mathematical skill practice to solve novel problems	0
Mixture of above categories	2
Not applicable	1

Qualitative comparative analysis

Due to the small number of maths app interventions available for this analysis (n = 8), the number of included app features within the QCA was restricted to three (see Table 8). Based on previous experimental research (e.g., Callaghan & Reich, 2021), the QCA model focused on the role of feedback (explanatory and motivational) and programmatic levelling (static or dynamic). A highly effective intervention was defined as a within-subject effect size (Cohen's *d*) greater than 1 (Hedge's *g* corrections applied to studies with sample sizes equal to or less than 50). In cases where math apps were evaluated in multiple studies (*onebillion Maths 3-5 and Maths 4-6; Math Shelf*), the most robust study (e.g., an RCT with the largest sample size) that had sufficiently reported data to calculate the within-subject effect size was used as an indication of children's learning outcomes (Outhwaite et al., 2018; Schacter & Jo, 2017).

All eight studies included in the QCA (Berkowitz et al., 2015; Cary et al., 2020; Grimes et al., 2020; Kosko & Ferdig, 2016; Outhwaite et al., 2018; Schacter & Jo, 2017; Spencer et al., 2013; Wu, 2020) focused on typically developing children aged between 4-7 years old with an experimental design (RCT or QED). Two studies reported sub-group analyses on children identified as underachieving in mathematics (Cary et al., 2020; Wu, 2020). But for the purposes of this analysis, the extracted within-group effect sizes focused on the whole sample.

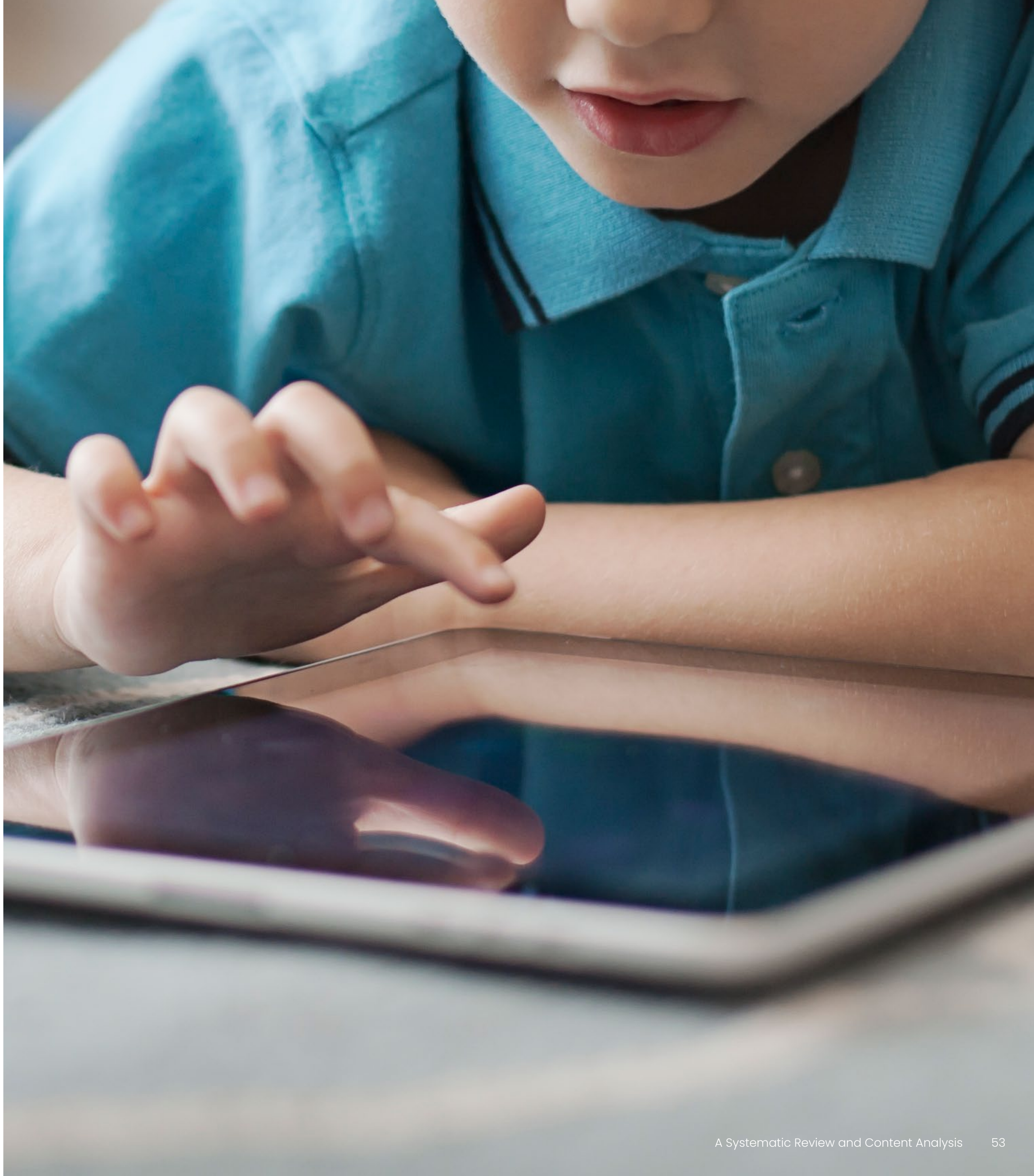


Table 8: Summary of included apps in the qualitative comparative analysis (QCA) with potential conditions and classifications of highly effective intervention sets.

Study	App	Potential Conditions					Outcome (Within-Group Effect Size)			Sample Size (Treatment Group only)
		Feedback-Explanatory	Feedback-Motivational	Levelling-Programmatic (Dynamic or Static)	Social Interaction	Task Instruction	Cohen's *d*	Hedge's g^a	Highly Effective Intervention Set^b	
Berkowitz et al. (2015)	*BedTime Math*	0	0	1	1	0	.82	n/a	0	Unknown
Cary et al. (2020)	*KinderTEK*	1	1	0	1	1	.88	n/a	0	58
Spencer et al. (2013)	*Know Number Lite*	0	1	0	0	0	.56	n/a	0	114
Schacter and Jo (2017)	*Math Shelf*	0	1	1	1	0	.20	n/a	0	231
Wu (2020)	*MathemAntics*	1	1	1	1	1	1.89	1.84	1	28
Grimes et al. (2020)	*Native Numbers*	1	1	1	0	1	1.10	1.06	1	24
Outhwaite et al. (2018)	*onebillion* *(Maths 3-5 and Maths 4-6)*	1	1	0	1	1	.78	n/a	0	126
Kosko and Ferdig (2016)	*Zorbit*	1	1	1	1	1	1.45	1.41	1	27

[a] *Hedge's g correction applied to effect sizes where study sample sizes were less than 50.*
[b] *Effect sizes greater than 1.00 coded as 1. Effect sizes less than 1.00 coded as 0.*

Table 9: Summary of Model 1 in the qualitative comparative analysis (QCA).

Feedback-Explanatory	Feedback-Motivational	Levelling-Programmatic (Dynamic or Static)	Number of Apps		Membership in 'Highly Effective Intervention' Set	Raw Consistency
1	1	1	3	*(MathemAntics; Native Numbers; Zorbit)*	1	1.00
1	1	0	2	*(KinderTEK; onebillion)*	0	n/a
0	1	1	1	*(Math Shelf)*	0	n/a
0	1	0	1	*(Know Number Lite)*	0	n/a
0	0	1	1	*(BedTime Math)*	0	n/a

Table 10: Summary of Model 2 in the qualitative comparative analysis (QCA).

Levelling-Programmatic Dynamic	Levelling-Programmatic Static	Feedback[c]	Number of Apps		Membership in 'Highly Effective Intervention' Set	Raw Consistency
1	0	1	2	*(Native Numbers; Zorbit)*	1	1.00
0	1	1	1	*(MathemAntics)*	1	1.00
0	0	1	2	*(KinderTEK; onebillion)*	0	n/a
1	0	0.5	1	*(Math Shelf)*	0	n/a
0	0	0.5	1	*(Know Number Lite)*	0	n/a
0	1	0	1	*(BedTime Math)*	0	n/a

Model 1

Results indicated the combination of explanatory and motivational feedback, together with programmatic levelling (static or dynamic, rather than participatory levelling) were necessary conditions for highly effective maths apps (see Table 9).

Although motivational feedback on its own was not associated with highly effective maths apps, importantly, it was not a hinderance on children's learning outcomes. It is also important to highlight that the differential effects of explanatory and motivational feedback could not be disentangled in this analysis. This is because within this sample, there were no maths apps that included explanatory feedback only.

Model 2

Model 1 focused on programmatic levelling, which can be dynamic or static. To further explore any differential effects between these different types of levelling, Model 2 included programmatic dynamic and programmatic static levelling entered as separate conditions, with feedback entered as a fuzzy set. In a fuzzy set, multiple categories can be included to represent a continuum (i.e., 0, 0.5, or 1), rather than the dichotomous variables (i.e., 0 or 1) used in Model 1. Within this fuzzy set, explanatory and motivational feedback was given a stronger weighting (1), compared to motivational feedback only (0.5) and no feedback (0; see Table 10).

Consistent with Model 1, results showed explanatory and motivational feedback combined with programmatic dynamic levelling were necessary conditions for highly effective maths apps. The same pattern of results was also observed for programmatic static levelling. This suggests there were no differential effects between the different types of programmatic levelling for enhancing children's learning outcomes within app-based mathematics instruction.

Discussion

The current content analysis identified that to date, practice-based maths apps are the most common type of app to be evaluated (n = 14). The included apps primarily targeted basic skills in understanding number representations and relationships (n = 21) with number representation with Arabic digits, verbal, and/or written number word recognition as the most common (n = 19). Most included apps incorporated some form of feedback (n = 21) and levelling (n = 21). Results from the QCA demonstrated that variations in how feedback and levelling was implemented with the app design was associated with differences in the within-subject effect sizes of children's progress following the maths app intervention.

Specifically, the QCA showed that the combination of explanatory and motivational feedback, together with programmatic levelling (either dynamic or static) were necessary app design features for enhancing children's learning outcomes with app-based maths instruction. These results are consistent with other experimental research that showed children made significantly fewer errors and completed app-based maths learning tasks quicker and

with increased accuracy with explanatory feedback and programmatic levelling (Callaghan & Reich, 2021).

Although the eight studies included in the QCA can be considered to have good levels of rigour using an experimental design (RCT or QED) with a relatively heterogenous sample of typically developing children, some caution should be taken with these results due to the possibility of inflated effect sizes, like the issues highlighted in Part 1. Although Hedge's *g* corrections were applied as appropriate to sample sizes less than 50, only two studies had overall final sample sizes over 250 children (Berkowitz et al., 2015; Outhwaite et al., 2018). In this QCA, both studies were not classified as a highly effective intervention set (within-subject effect size greater than 1), relative to the other included studies. Likewise, of the three studies that were identified as a highly effective intervention set (Grimes et al., 2020; Kosko & Ferdig, 2016; Wu, 2020), two used a researcher developed assessment of mathematical abilities as the primary outcome measure. Within this QCA, these issues cannot be statistically controlled for, and so instead should be considered as potential caveats when interpreting the study results. Nevertheless, these results can be used to inform the development of clear hypotheses in future research on how educational maths apps work to support learning with young children.

Application to the Top 25 Most Popular Commercial Educational Maths Apps

The coding framework developed for the content analysis (Part 2) was also applied to the Top 25 most popular commercial educational maths apps from the iOS Apple App and Google Play Stores (see Appendix 5). The term "maths" was used to search the Apple App and Google Play stores on 11th and 12th August 2021 and results were sorted by popularity. The Top 25 most popular commercial educational maths apps for 5-year-old children present in both app stores were collected. The search strategy excluded app bundles from the count; only singular apps were counted as part of the Top 25.

Of these 25 maths apps identified, 18 included some form of mathematical content, and one app has been empirically evaluated (*onebillion Maths 4-6*; see Appendix 1). Six apps did not include any mathematical content (see Appendix 6). Most of the 18 commercial educational maths apps were also practice based (n = 17) and targeted number (n = 16) and counting skills (n = 16), mostly in isolation (n = 17). Most apps included motivational feedback (n = 14) and participatory, free-form levelling opportunities within their design (n = 17), as well as social interaction, typically through the form of an in-app character (n = 18). However, very few gave task instructions (n = 4).

Overall, these results demonstrate that many of the commercial educational apps for young children that are categorised as 'maths', are not necessarily reflective of best practices in app content and design.

Most apps did not comprehensively capture all areas of mathematical development, nor did they adequately include features of personalisation, such as explanatory feedback and programmatic personalisation, which this research has shown maximises children's outcomes in app-based learning. This demonstrates the limited options for identifying high-quality maths apps currently available for parents and teachers and highlights the need to improve the meaningful categorisation of educational apps on the app stores to facilitate parent and teacher choice.

Overall Conclusions

Overall, the current study has conducted a comprehensive systematic review and synthesised the wide range of current evidence on educational maths apps for young children in the first three years of compulsory school (Part 1). The inclusion of quantitative, qualitative, and mixed methods studies has enabled a diverse set of questions to be addressed focused on 'what works', including do educational maths apps work, how do they work, under what circumstances to do they work, and for whom do they work for best? While the current evidence base demonstrates promising findings for supporting children's outcomes, there is a clear need for more high-quality, rigorous evaluation studies that are guided by the current findings and gaps highlighted in this review.

The current study has also focused more in-depth on how educational maths apps work to support learning through a content analysis on the mathematical content and design features included within the identified maths apps (Part 2). This has identified feedback and levelling as key mechanisms for supporting young children in an app-based learning context, as well as also highlighting important directions for future research to build on this evidence base.

Considering the current evidence, it can be concluded that educational maths apps can, in some cases, support the mathematical learning outcomes of typically developing children, when effectively implemented in a classroom setting. When deciding if and which educational maths apps to use with children, teachers, parents, and policy makers should recognise that technology alone will not equal success.

To achieve the best outcomes, parents and teachers should consider: 1) whether children can access the chosen maths app based on their individual skills, particularly relating to their language and existing mathematical attainment. 2) How the chosen maths apps can be meaningfully integrated into a well-rounded maths curriculum and implemented with appropriate support systems for a sustained period. 3) Whether the chosen apps contain explanatory feedback, whereby the app explains why the answer is right or wrong and programmatic levelling, whereby the learning content is scaffolded and personalised to the individual child. Furthermore, policy makers should also consider how the apps can be accessed, selected, and used by schools and families at home if they choose. In the UK, poverty is rising fastest among families of young children (Department for Work and Pensions, 2019).

Addressing digital exclusion is a key consideration in ensuring equitable access to digital learning opportunities for young children and online guidance for parents, particularly in the context of the home learning environment. These recommendations will help guide the translation of these current research findings into practice and ensure that maths apps can add value to young children's learning.

Additional Resources

All resources and data for this project can be found on the Open Science Framework project page **(https://osf.io/pzkmh/?view_only=04aaebc9456143119e420fcff01e7306/).**

Appendix

Appendix 1: Descriptive summaries of the 50 included studies in the systematic review (Part 1)

Ahmad et al. (2014) observed five 9-year-old children with Down syndrome using *MathDS* in Malaysia (app available in English and Malay). The observations were conducted as part of initial user acceptance testing in the development of the app. After the observed session, four children demonstrated improvements in recalling and writing numbers one to five. One child displayed no interest and was disengaged with the app.

Berggren and Hedler (2014) conducted a qualitative exploration with 30 4-5-year-old children in Sweden to examine their enjoyment of using the *CamQuest* app. The authors found that children enjoyed using the app which allowed practice in collaboration, communication, and problem solving.

Berkowitz et al. (2015) conducted a pupil-level RCT evaluating *Bedtime Math* compared to the reading app, Bedtime Learning Together in the USA. Importantly, the design and implementation of the apps were consistent in both groups; only the educational content was changed. 278 6-7-year-old children and their caregivers used the apps, on average, once a week for 22 weeks. Results showed stronger learning gains in mathematics for the maths app (Cohen's *d* = .82), compared to the reading app. The maths app intervention was also found to be particularly beneficial for children of parents with high maths anxiety.

Broda et al. (2019) conducted a single case design study with 18 4-5-year-old children in the USA to examine subitising speed and accuracy following the use of *Fingu* for one month. In app data were used and highlighted that over the course of the intervention, children became faster and more accurate on tasks. Girls were reported to be more accurate in the subitising tasks of the app than boys.

Bullock et al. (2017) assessed the feasibility of three maths apps (*Intro to Math; Montessori Pink Tower; Montessori Numbers for Kids Base-10 Blocks*) with 19 4-year-old children in the USA. Children also completed a one-to-one clinical interview, but in contrast to the other studies evaluating these apps (Moyer-Packenham et al., 2016; Watts et al., 2016), results showed most children did not make any significant progress in seriation and counting skills.

Cary et al. (2020) compared different dosage of *KinderTEK* to standard practice**,** with 114 5-6-year-old children in the USA. Overall, results showed children who used the maths app made significant improvements in mathematics (Cohen's *d* = .88) compared to a business-as-usual control group. When comparing dosage, children in the Early Start group, who used the maths app for three times a week across the school year (three terms), made stronger learning gains compared to the Late Start group, who also used the maths app for three times a week, but for only two school terms.

Cornu et al. (2019) found the maths app *MaGrid*, significantly improved 5-6-year-old children's visuo-spatial skills in Luxembourg (Cohen's *d* = .71 [averaged across the two sub-tests]), targeted by the maths app intervention, relative to controls (n = 125). However, the training effects did not transfer to broader mathematical skills.

Ginsburg et al. (2019) presented a qualitative case study of one 4-year-old child in the USA. The child used the *MathemAntics* app on one occasion (24 minutes) with the interviewer (second author). The authors reported that the child enjoyed using the app, particularly the visual and audio features.

Griffith et al. (2019) evaluated a selection of 18 educational apps, of which seven targeted early mathematical skills (*Bugs and Numbers; Counting Caterpillar; Grover's Number Special; Moose Math; Motion Math: Hungry Guppy; Park Math; Tally Tots*), compared to 18 entertainment apps with no mathematical content using a pupil-level RCT design in the USA with 22 4-5-year-old children. Results showed children who used the maths apps for three months made significantly greater improvements in mathematical skills (Cohen's *d* = .87), compared to the active control group. However, as the results are not disaggregated per app, it was not possible to identify the specific benefits of each app evaluated.

Grimes et al. (2020) compared the maths app *Native Numbers* to supplementary one-to-one and small peer group maths (and other subjects) instruction in the USA with 46 5-6-year-old children. Results showed children made significant gains in number sense skills following use of the maths app intervention (Cohen's *d* = 1.10), compared to the active control group. However, no significant improvements were reported for mathematical language skills.

Hasanah et al. (2017) evaluated the *Monsakun* app through a single case design study with 39 children aged 6-7 years in Japan. The study aimed to examine in-app behaviour through log data, including the average number of steps made by children and the number of errors made. Results highlighted that *Monsakun* was beneficial for learning arithmetic word problems of one-step addition and subtraction. The authors found that despite some children taking more steps in some in-app tasks and posing incorrect arithmetic word problems, the answers were mostly meaningful, and many children answered correctly.

Hassler Hallstedt et al. (2018) conducted a pupil-level RCT with four treatment and control groups to examine the impact of *Chasing Planets* with 281 8-9-year-old children underachieving in maths in Sweden. Children were randomly assigned to: a) use *Chasing Planets* only, b) use *Chasing Planets* and an app-based working memory game, c) use an adapted version of *Chasing Planets* focused on reading skills (active control), or d) standard practice (business-as-usual control). Results showed children in both groups using the maths app intervention for 19-20 weeks had significantly greater gains in basic arithmetic skills (Cohen's *d* = 1.19), compared to the two control groups. However, no arithmetic transfer or problem-solving effects were observed across any of the four groups.

Hieftje et al. (2017) found 5-6-year-old children in the USA who used the maths app *Knowledge Battle* made significant improvements in maths performance, but only gains in the numeration sub-test were significantly greater than children using an app-based attention/time control game, with no maths content. No difference was observed in other mathematical skills, including measurement, numerical operations, and problem solving.

Hung et al. (2015) implemented a QED with 43 7-8-year-old children in Taiwan to examine the app *Motion Math: Hungry Fish*. The intervention group played levels 7-14 in *Motion Math: Hungry Fish*, whilst the control group played levels 1-6. The authors reported positive results in favour of the maths app intervention for learning outcomes and flow experience, compared to the active control group. However, there were no differences in the intrinsic motivation across the intervention and control group.

Judd and Klingberg (2021) conducted a pupil-level RCT to evaluate different versions of the *Vektor* app with 6-8-year-olds in Sweden (n = 17,648). Results showed that training mathematical reasoning and visuo-spatial working memory had the most impact, compared to other components of visual spatial skills.

Kalmpourtzis (2014) evaluated the feasibility of *LadyBug Box* using structured observations with 17 4-5-year-old children in Greece. Results showed children were able to effectively use the maths app independently and collaboratively, and demonstrated improvements in spatial thinking, navigation, and mental representation skills.

Kosko and Ferdig (2016) conducted a RCT with 50 children aged 4-5 years in the USA to evaluate the impact of *Zorbit's Math Adventure for Preschool* on maths attainment. The intervention was home-based, with parents asked to ensure children in the intervention group played with the app on a weekly basis. Positive results in favour of the intervention on maths attainment compared to standard mathematical practice in the home (control group) were reported.

Kromminga and Codding (2020) used an adapted alternating treatment design (within-subject) to evaluate *Quizlet Plus* in the USA. Teacher reports were used to identify four 7-8-year-old children in need of additional support in mathematics. Children were then also screened by the research team to identify strengths and weaknesses in addition and subtraction skills. This supplementary assessment tool was used to target the intervention to the children's individual learning needs. Results showed children made significant gains in maths skills with both the app- and paper-based versions of intervention.

Lee and Choi (2020) found 6-10-year-old children in Tanzania made significant improvements in number identification, quantity discrimination, addition, subtraction, and missing number skills (Cohen's *d* = .60 [averaged across the five sub-tests]) after using *KitKit School* (n = 61). The strongest gains were seen in number identification skills, and analysis of children's gameplay data showed tasks relating to these skills, including recognising numbers with objects, number writing, and counting objects, were the most frequently played. However, the gains in number identification and missing number skills were not statistically significant from that of the business-as-usual control group.

Litster et al. (2019) conducted one-to-one clinical interviews with 65 5-8-year-old children in the USA to explore engagement with the app *Montessori Number Base-10 Blocks.* The authors recorded the interviews and qualitatively analysed the video data. Results indicated mixed engagement behaviour when observing young children using the app.

Mattoon et al. (2015) found 4-5-year-old children (n = 24) in the USA using five maths apps (*Motion Math: Hungry Fish; DinoKids-Math Lite; Math Express; onebillion Maths 3-5; Park Math*) made significant improvements in maths performance (Cohen's *d* = 1.09), but these gains were not statistically significant compared to children in the active control group using traditional manipulatives, such as plastic bears, Unifix cubes, and dice.

Miller (2018) included a traditional play-based learning active control group in their study of 13 4-5-year-old children in Canada. Results showed after the 2-week intervention, children who used a selection of 15 math apps (*Addition; Basic Skills; Candy Count; Colouring Book; Count-up-to-ten; Endless 123; Intro to Math; Junior Math; Kindergarten; Lola's Math; Math Game; Math School; Preschool all-in-one; Preschool and Kindergarten Learning Games; Splash Math Pre-school and Kindergarten*) had marginally higher, but not statistically significant gains in number, counting, and measurement skills (Cohen's *d* = .05), compared to the active control group. However, while some guidance on which apps to use was given to the treatment group, the frequency and duration of use for each app were not clear and the results per app were not disaggregated. Miller (2018) also conducted interviews with teachers and observations of children's interactions with the 15 maths apps. Qualitative analyses identified that children often collaborated with each other and preferred constructive apps (i.e., virtual manipulatives) over practice-based apps with appropriate levels of difficulty and sequencing of mathematical content. Children often became disengaged with the apps when the content became too difficult for their attainment level or attention span.

Moyer-Packenham et al. (2016) conducted feasibility evaluations of 11 maths apps (*100s Board; Fingu; Friends of Ten; Intro to Math; Montessori Numbers for Kids Base-10 Blocks; Montessori Number for Kids Place Value Cards; Montessori Numbers for Kids Skip Counting Beads; Montessori Pink Tower; Motion Math: Hungry Guppy; Motion Math: Zoom; Teaching Number Lines*) with 65 5-8-year-old children in the USA. Children completed a one-to-one clinical interview and comparisons of pre- and post-test performances showed their efficiency in quantity and skip counting, and accurate performance in subitising and skip counting significantly improved over time.

Nunes et al. (2019) conducted a large-scale, pupil-level RCT efficacy trial of the *onebillion Maths 3-5 and Maths 4-6* apps with children identified as underachieving in mathematics by their teachers across 114 schools in England. In a final sample of 1,089 5-6-year-old children, results indicated low achieving children using the maths app intervention for 12 weeks showed significantly greater improvements, compared to a business-as-usual control group of children also identified as underachieving in mathematics.

Outhwaite et al. (2017) assessed the feasibility of the *onebillion Maths 3-5, Maths 4-6, Count to 10, and Count to 20* apps with 4-5-year-old children in England with four studies. Study 1 (n = 26) showed children made significant learning gains in curriculum knowledge (Cohen's *d* = 1.01), which generalised to conceptual maths skills (Cohen's *d* = .31) after using the apps for six weeks. The observed learning gains were sustained for curriculum knowledge six months later. Study 2 (n = 18) and Study 3 (n = 27) replicated these immediate learning gains with different samples of children over a longer 13-week intervention duration. Study 4 (n = 27) examined whether the maths app intervention was beneficial for low-achieving children (n = 12) compared to the business-as-usual control group (n = 15; relatively high-achieving age-matched controls from the same class). Results highlighted a positive effect in favour of the maths app intervention after eight weeks compared to the business-as-usual control group. More specifically, the intervention group made significantly greater improvements in maths curriculum knowledge than the control group.

Outhwaite et al. (2018) conducted a pupil-level RCT to evaluate *onebillion Maths 3-5 and Maths 4-6* with 389 4-5-year-old children in England. Results showed following the 12-week intervention, children made significantly greater improvements in basic and higher-order mathematical skills when using the maths apps instead of a daily small-group maths activity (time-equivalent treatment, Cohen's *d* = .65) or in addition to standard practice (treatment, Cohen's *d* = .78), compared to standard practice only (business-as-usual control). When disentangling the effects of the maths app intervention as a form of quality instruction (time-equivalent treatment) from additional time spent learning maths (treatment), no significant differences were observed, suggesting the apps could be implemented as part of a well-balanced curriculum and not take away from other important subjects.

Outhwaite et al. (2019) conducted a mixed methods IPE to understand how contextual factors within the learning environment influenced learning outcomes reported in Outhwaite et al. (2018). Observations of the maths app intervention sessions and interviews with the participating teachers identified four implementation themes: teacher support, teacher supervision, intended implementation, and established routine. A regression analysis showed 41% of the variance in children's learning outcomes were best explained by the established routine theme, which included implementing the intervention at a consistent time each day, having a member of staff responsible for the intervention implementation, having well-organised equipment, and having dedicated space within the classroom and a seating plan for the maths app intervention.

Outhwaite et al. (2020) compared the effects of implementing *onebillion Maths 3-5 and Maths 4-6*, in different languages with 61 5-6-year-old bilingual children in Brazil. Children who used the maths apps for 10 weeks made significantly greater gains in mathematics (first language Cohen's *d* = 1.46; second language Cohen's *d* = 1.06), compared to the business-as-usual control group who received standard practice (delivered in English). However, children who used the maths apps in their first language (Brazilian Portuguese) made greater progress through the app (i.e., more topics completed) than those who used the apps in their second language (English).

Parks and Tortorelli (2020) compared the effects of nine math apps (*3-2-1 Snack; Bubble Bath; Chicken Coop; Happy Camel; Mega Mall; Paint-A-Long; Pan Balance; Peg's Pizza Place; Race Car*) to an active control group with 298 5–6-year-old children in the USA, including five other educational apps covering literacy, science, and maths. Following the 7–8-month intervention period, results showed no significant differences in maths gains between the two groups. However, the effects of each app were not disaggregated and only 7.8% of the treatment group met the recommended time of playing apps for at least 30 minutes per week.

Pecora (2015) conducted structured observations and pre- and post-test mathematics assessments to evaluate *GoMath!* with six 5-6-year-old children with special educational needs (ADHD and communication impairments) in the USA. Children used the app in addition to standard practice for six weeks and showed significant improvements in mathematical skills. However, the structured observations conducted by the research-teacher indicated children were only focused for 28% of the overall instructional time. Children were frequently distracted (36% of the time), unfocused (24% of the time), or affected by schedule changes (12% of the time). As such, the author highlighted the behaviour of children and classroom environment may have affected app engagement and emphasised challenges of implementing app-based math instruction with children with special educational needs.

Pires et al. (2019) evaluated two different implementations of the maths app *BrUNO* with 60 6-7-year-old children in Uruguay, compared to a business-as-usual control group. Following the 3-week intervention period, results showed children who used the *BrUNO* app with a set of physical play blocks of different sizes (similar to Cuisenaire rods) made the strongest learning gains (Cohen's *d* = .40), compared to using the same *BrUNO* app with virtual blocks (Cohen's *d* = .35) and standard mathematical practice. Recordings of the children's interactions with the maths app intervention showed children used a greater number of physical blocks and more efficient task completion strategies, compared to the virtual manipulatives.

Pitchford (2015) also conducted a pupil-level RCT evaluating *onebillion Maths 3-5, Maths 4-6, Count to 10, and Count to 20* (in Chichewa) with 283 children in the first three years of primary school in Malawi. Due to the ability-based educational system in Malawi the age of participating children ranged from 6 to 13 years. Following the 8-week intervention, children who used the maths app intervention made significantly greater learning gains in curriculum knowledge (Cohen's *d* = 1.80) and transferred to conceptual maths skills (Cohen's *d* = .88), compared to children using other apps with no educational content (active control) and standard practice (business-as-usual control).

Pitchford et al. (2018) conducted structured video observations of 32 7-11-year-old children with vision loss, and emotional, behavioural, communication and learning difficulties in Malawi using *onebillion Maths 3-5 and Maths 4-6* (in Chichewa). The qualitative data was combined with quantitative monitoring data automatically collected on children's progress rate through the app. Results showed children with special educational needs and disabilities were able to interact with the app-based math instruction, but their average progress rate was twice as long relative to their mainstream peers. The extent of children's additional needs also significantly predicted progress rate; children with hearing and/or language difficulties made slower progress compared to other special educational needs and disability profiles.

Pitchford et al. (2019) implemented a QED with 256 5–11-year-old children in Malawi to evaluate *onebillion Maths 3-5 and Maths 4-6* (in Chichewa). Across the 14-month intervention period, the intervention group interacted with the apps for a total of 540 minutes. The business-as-usual control group received standard mathematics classroom instruction delivered by teachers. Positive results in favour of the maths app intervention were reported compared to standard mathematical practice (control group). In addition, the authors reported that the maths app intervention could effectively support the mathematical development of girls in contexts, such as Malawi, where gender differences in standard practice may hinder their learning progress.

Ramani et al. (2019) also conducted a pupil-level RCT evaluating *The Great Race,* with 148 5-6-year-old children in the USA. Children who used the maths app for ten sessions showed greater learning gains in numerical knowledge at the latent level (Cohen's *d* = .14), compared to children using an app with the same layout as The *Great Race,* but with no mathematical content (active control) or the working memory app *Recall them All* (active control).

Schacter et al. (2016) used a pupil-level RCT design to compare the impact of *Math Shelf* to the five most downloaded and best reviewed maths apps of 2014 with 86 children aged 4-5 years old in the USA. Following the 6-week intervention, children from low-income families who used the *Math Shelf* app made significantly greater improvements in their number sense knowledge (Cohen's *d* = .90), compared to their peers in the active control group.

Schacter and Jo (2016) implemented a QED in the USA with 162 4–5-year-old children. The experiment evaluated the app *Math Shelf* across a 15-week intervention period. The intervention group interacted with the app for a total of 300 minutes across the intervention period, while a business-as-usual control group continued to receive standard mathematics classroom instruction delivered by teachers. The study reported positive results in favour of the maths app intervention compared to standard mathematical practice. The authors concluded that children in the intervention group learned approximately one year more mathematics than children in the control group.

Schacter and Jo (2017) conducted a school-level RCT of 378 4-5-year-old children in the USA. Results showed children who used the *Math Shelf* app for 22 weeks made significantly greater improvements in mathematical skills (Cohen's *d* = .20), compared to their peers receiving research-based, hands-on maths lessons covering the same mathematical content.

Schaeffer et al. (2018) conducted a 2-year follow up of the same cohort of participants in Berkowitz et al. (2015) using the *Bedtime Math* app compared to the reading app, *Bedtime Learning Together*. Results showed that although app usage significantly decreased over time, the initial learning gains on the same standardised mathematical outcome measure were sustained.

Schenke et al. (2020) conducted a pupil-level RCT with three treatment and control groups to evaluate *MeasureUp!* with 99 4-5-year-old children in the USA. Results showed following the 3-week intervention, children made significantly greater improvements in mathematical skills with the *MeasureUp!* app (Cohen's *d* = .65) and in combination with *SuperVision,* a companion app for parents (Cohen's *d* = .53), compared to an active control group using the literacy app *SuperWHY!* However, there were no significant differences between the two forms of app implementation. Importantly, children used the maths app in school and although parents were involved through the companion app, the authors concluded that more consideration is needed for designing effective parental supports.

Spencer (2013) implemented a QED in Dubai, United Arab Emirates, with 114 4–5-year-old children to evaluate *Know Number Lite* on numeracy learning. The intervention group interacted with the app for two 10-minute sessions a day, for five consecutive days. The business-as-usual control group continued to receive standard mathematics classroom instruction delivered by teachers. The study found positive results in favour of the maths app intervention compared to standard mathematical practice. However, no significant impact of the maths app intervention on children's motivation was found when compared to the control group.

Stacy et al. (2017) conducted qualitative observations, interviews, and field notes within a community-based participatory research approach to examine the feasibility of the *IXL* maths app with young children in the USA. Results showed children responded well to the maths app intervention and minimal efforts were required from the supervising adults, who reported a lack of expertise in maths education, as well as their own maths anxieties and struggles. The rich, qualitative data also identified barriers to the maths app intervention focused on the costs of the technology, reliable internet connections, and how to promote long-term adherence and motivation.

Stubbé et al. (2016) implemented a QED with 703 children aged 7-9 years in Sudan using *E-learning Sudan.* The intervention group accessed the app for a period of six months, with a total time on task of 90 hours (5,400 minutes). The control group continued to receive standard mathematical practice taught by teachers. The authors reported positive results in favour of the maths app intervention compared to standard mathematical practice.

Swicegood (2015) found no significant improvements in mathematical skills with 40 children aged 6-7 years in the USA using the *Addimal Adventure and Splash Math 2nd Grade* apps in addition to standard practice for four months. Observations showed children did not effectively engage with the maths app intervention. Children frequently tapped the screen until the correct answer was selected, which was thought to be a consequence of the game not providing feedback on incorrect answers. Interview data also showed that although children reported more confidence using the maths app intervention, there was a decrease in affective engagement, which was suggested to underpin the lack of observed improvement in mathematical learning outcomes.

Tucker et al. (2016) used qualitative methods to identify a taxonomy capturing the complex and multidimensional relationships between 7–8-year-old children's access to app affordances and how this supported or hindered their learning outcomes with six maths apps (*100s Board; Montessori Numbers for Kids Base-10 Blocks; Montessori Numbers for Kids Bead Skip Counting; Montessori Numbers for Kids Place Value; Motion Math*: *Zoom; Teaching Number Lines*).

Vanbecelaere et al. (2020) found the *Number Sense Game*, significantly improved 6-7-year-old children's targeted number line estimation performance (Cohen's *d* = 1.58), in comparison to controls in Belgium (n = 222). However, no significant group differences were observed for digit comparison skills, which were also targeted by the maths app intervention. Training effects did not transfer to overall mathematical competence at immediate or delayed post-test, and no benefits were found for children's maths anxiety.

Watts et al. (2016) conducted further micro-scoring analyses of the 11 maths apps explored by Moyer-Packenham et al. (2016). The sample of 65 5-8-year-old children in the USA was the same as those participating in the studies by Litster et al. (2019) and Moyer-Packenham et al. (2016). This study also demonstrated that although not all areas of mathematics assessed showed statistically significant increases in speed and accuracy, most children demonstrated descriptive, incremental shifts in their ability to subitise, understand quantity, place value, and skip counting.

Wu (2020) evaluated *MathemAntics* through a RCT with 56 children aged 4-5 years in the USA. The intervention group interacted with the app for 15 minutes a day, three days a week for four weeks (overall time on task was 180 minutes). The business-as-usual control group continued to receive standard mathematical practice in the classroom. Wu (2020) reported positive results in favour of the maths app intervention compared to the control group. The study also included a focus on children identified as underachieving in mathematics through sub-sample analyses. The maths app intervention had a positive impact on the mathematical learning of children identified as underachieving (Cohen's d = 2.11), when compared to their higher-attaining peers.

Zander et al. (2016) evaluated *Rotate It!* in comparison to a paper version of the same spatial rotation task. Thirty-seven children aged 8-11 years in Germany worked with both conditions in different, counterbalanced orders. Results showed an additive effect of the app-based instruction, in which children demonstrated significant improvements in task solving performance when first completing the paper-based static version of the task, followed by the app-based, dynamic version. Training effects were not observed in the opposite form of implementation.

Zhang et al. (2020) found 8-9-year-old children in China using the apps *Motion Math* and *Slice Fractions*, made significant improvements in fraction skills (Cohen's d = 2.31), but these gains were not significantly different from the business-as-usual control group, who received the same duration of mathematical instruction (n = 65). However, children in the intervention group demonstrated stronger performance in the transfer test of magnitude comparison, relative to controls.

Appendix 2: Summary of 50 included studies in the systematic review (Part 1)

Study	Country, Setting (Implementer)	Final Sample (n, age)	App(s)	Method	Controls	Primary Maths Outcome Measure	Time on Task (minutes)	Overall results	Primary Effect Size*	Quality Score	Attrition Rate %
Ahmad et al. (2014)	Malaysia, School (NR)	SEND (5, 9 years)	*MathDS*	Qualitative	No control	Researcher developed	NR	Mixed	NR	.55	0
Berggren & Hedler (2014)	Sweden, School (Practitioner)	TD (30, 4-5 years)	*CamQuest*	Qualitative	No control	NA- enjoyment only	180	Positive towards app	NA	.45	0
Berkowitz et al. (2015)	USA, Home (Parent)	TD (278, 6-7 years)	*BedTime Math*	RCT	Active- other educational apps	Standardised- Woodcock-Johnson-III	NR	Positive towards app	.82	.89	52.6
Broda et al. (2019)	USA, School (Practitioner and researcher)	TD (18, 4-5 years)	*Fingu*	Single case design	No control	In-app data	400	Positive towards app	NA	.64	0
Bullock et al. (2017)	USA, School (Researcher)	TD (19, 4-5 years)	*3 apps*	Mixed methods	No control	Researcher developed	30-40	Mixed	NA	.67	0
Cary et al. (2020)	USA, School (Practitioner)	TD/ LA (114, 5-6 years)	*KinderTEK*	QED	Multiple	Standardised- ASPENS	298	Positive towards app	.88	.68	11.6
Cornu et al. (2019)	Luxembourg School (Researcher)	TD (125, 5-6 years)	*MaGrid*	RCT	BAU	Researcher developed	400	Mixed	.71	.75	4.6
Ginsburg et al. (2019)	USA, School (Researcher)	TD (1, 4 years)	*MathemAntics*	Qualitative	No control	Researcher developed	24	Positive towards app	NA	.60	0

Appendix 2: Summary of 50 included studies in the systematic review (Part 1)

Study	Country, Setting (Implementer)	Final Sample (n, age)	App(s)	Method	Controls	Primary Maths Outcome Measure	Time on Task (minutes)	Overall results	Primary Effect Size*	Quality Score	Attrition Rate %
Griffith et al. (2019)	USA, Home (Parent)	TD (22, 4-5 years)	*7 apps*	RCT	Active- placebo	Standardised - TEMA-3	NR	Positive towards app	.87(.84)	.79	0
Grimes et al. (2020)	USA, School (Practitioner)	TD (46, 5-6 years)	*Native Numbers*	RCT	Active – non-digital	Standardised-Number Sense Screener	660	Mixed	1.10 (1.08)	.79	0
Hasanah et al. (2017)	Japan, School (Practitioner)	TD (39, 6-7 years)	*Monsakun*	Single case study	No control	In-app data	70	Positive towards app	NA	.46	0
Hassler Hallstedt et al. (2018)	Sweden, School (Practitioner)	LA (281, 8-9 years)	*Chasing Planets*	RCT	Multiple	Standardised - Math Battery	1,122- 1,698	Mixed	1.19	.93	.7
Hieftje et al. (2017)	USA, School (NR)	TD/LA (133, 5-6 years)	*Knowledge Battle*	QED	Active- placebo	Standardised-KeyMath-3	480-720	Mixed	NR	.61	.7
Hung et al. (2015)	Taiwan, School (Practitioner)	TD (43, 7-8 years)	*Motion Math: Hungry Fish*	QED	Active- other educational apps	Researcher developed	NR	Positive towards app	NR	.50	17.3
Judd & Klingberg (2021)	Sweden, School (Practitioner)	TD (17,648, 6-8 years)	*Vektor*	RCT	Active- other educational apps	Researcher developed/ in-app data	720-1,188	Positive towards app	NR	.68	59.0
Kalmpourtzis (2014)	Greece, School (Practitioner)	TD (17, 4-5 years)	*LadyBug Box*	Qualitative	No control	Researcher developed	NR	Positive towards app	NR	.60	0
Kosko & Ferdig (2016)	USA, Home (Parent)	TD (50, 4-5 years)	*Zorbit's Math Adventure for Preschool*	RCT	BAU	Researcher developed	NR	Positive towards app	1.45 (1.43)	.61	0
Kromminga & Codding (2020)	USA, School (Researcher)	LA (4, 7-8 years)	*Quizlet Plus*	Adapted alternating treatment design	Multiple	Researcher developed	405- 510	Mixed	NR	.61	0

Appendix 2: Summary of 50 included studies in the systematic review (Part 1)

Study	Country, Setting (Implementer)	Final Sample (n, age)	App(s)	Method	Controls	Primary Maths Outcome Measure	Time on Task (minutes)	Overall results	Primary Effect Size*	Quality Score	Attrition Rate %
Lee & Choi (2020)	Tanzania, School (Researcher)	TD (61, 6-10 years)	*KitKit School*	RCT	BAU	Researcher developed	NR	Mixed	.60	.68	50.0
Litster et al. (2019)	USA, School (Researcher)	TD (65, 5-8 years)	*Montessori Numbers for Kids Base-10 Blocks*	Qualitative	No control	NA-engagement only	30-40	Mixed	NA	.90	0
Mattoon et al. (2015)	USA, School (Practitioner)	TD (24, 4-5 years)	*5 apps*	QED	Active – non-digital	Standardised-TEMA-3	180	No difference	1.09 (1.05)	.68	0
Miller (2018)	Canada, School (Researcher)	TD (13, 4-5 years)	*15 apps*	RCT + IPE	Active – non-digital	Researcher developed	200	No difference	.05 (.05)	.72	0
Moyer-Packenham et al. (2016)	USA, School (Researcher)	see Litster et al. (2019)	*11 apps*	Mixed methods	No control	Researcher developed	see Litster et al. (2019)	Mixed	NA	.90	NA
Nunes et al. (2019)	England, School (Practitioner)	LA (1,089, 5-6 years)	*onebillion Maths 3-5 and Maths 4-6*	RCT + IPE	BAU	Standardised-PTM5/6	1,440	Positive towards app	NR	.87	3.1
Outhwaite et al. (2017) Study 1	England, School (Practitioner)	TD (26, 4-5 years)	*onebillion Maths 3-5 and Maths 4-6. Count to 10 and Count to 20*	QED	No control	Researcher developed	900	Positive towards app	1.01 (.98)	.71	0
Outhwaite et al. (2017) Study 2	England, School (Practitioner)	TD (26, 4-5 years)	*onebillion Maths 3-5 and Maths 4-6.*	QED	No control	Researcher developed	1,950	Positive towards app	1.32 (1.26)	See Study 1	0
Outhwaite et al. (2017) Study 3	England, School (Practitioner)	TD (26, 4-5 years)	*onebillion Maths 3-5 and Maths 4-6.*	QED	No control	Researcher developed	1,950	Positive towards app	1.81 (1.76)	See Study 1	0
Outhwaite et al. (2017) Study 4	England, School (Practitioner)	TD (26, 4-5 years)	*onebillion Maths 3-5 and Maths 4-6.*	QED	BAU	Researcher developed	1,200	Positive towards app	3.34 (3.24)	See Study 1	0

Appendix 2: Summary of 50 included studies in the systematic review (Part 1)

Study	Country, Setting (Implementer)	Final Sample (n, age)	App(s)	Method	Controls	Primary Maths Outcome Measure	Time on Task (minutes)	Overall results	Primary Effect Size*	Quality Score	Attrition Rate %
Outhwaite et al. (2018)	England, School (Practitioner)	TD/LA (389, 4-5 years)	*onebillion Maths 3-5 and Maths 4-6*	RCT	Multiple	Standardised-PTM5	1,800	Positive towards app	.78	.86	15.6
Outhwaite et al. (2019)	England, School (Practitioner)	see Outhwaite et al. (2018)	*onebillion Maths 3-5 and Maths 4-6*	IPE	see Outhwaite et al. (2018)	NA- implementation only	NA	NA	NA	.85	NA
Outhwaite et al. (2020)	Brazil, School (Practitioner)	TD (61, 5-6 years)	*onebillion Maths 3-5 and Maths 4-6*	QED	Multiple	Standardised-EGMA	800	Positive towards app	1.46	.71	1.6
Parks & Tortorelli (2020)	USA, School (Practitioner)	TD (298, 5-6 years)	*9 apps*	RCT + IPE	Active- other educational apps	Standardised-AIMSweb test of Early Numeracy	600-1,350	No difference	NR	.75	43.3
Pecora (2015)	USA, School (Practitioner)	SEND (6, 5-6 years)	*GoMath!*	Mixed methods	No control	Researcher developed	NR	Mixed	NR	.70	0
Pires et al. (2019)	Uruguay, School (Researcher)	TD (60, 6-7 years)	*BrUNO*	QED	Multiple	Standardised-TEMA-3	260	Mixed	.40	.64	6.3
Pitchford (2015)	Malawi, School (Practitioner)	TD (283, 6-13 years)	*onebillion Maths 3-5 and Maths 4-6. Count to 10 and Count to 20*	RCT	Multiple	Researcher developed	600-1,200	Positive towards app	1.80	.71	11.0
Pitchford et al. (2018)	Malawi, School (Practitioner)	SEND (32, 7-11 years)	*onebillion Maths 3-5 and Maths 4-6*	Mixed methods	No control	In-app data	NR	Positive towards app	NA	.90	3.0
Pitchford et al. (2019)	Malawi, School (Practitioner)	TD (256, 5-11 years)	*onebillion Maths 3-5 and Maths 4-6*	QED	BAU	Standardised-EGMA	540	Positive towards app	NR	.89	79.0

Appendix 2: Summary of 50 included studies in the systematic review (Part 1)

Study	Country, Setting (Implementer)	Final Sample (n, age)	App(s)	Method	Controls	Primary Maths Outcome Measure	Time on Task (minutes)	Overall results	Primary Effect Size*	Quality Score	Attrition Rate %
Ramani et al. (2019)	USA, School (NR)	TD (148, 5-6 years)	*The Great Race*	RCT	Multiple	Researcher developed	100-150	Positive towards app	.14	.71	9.2
Schacter et al. (2016)	USA, School (Researcher)	TD (86, 4-5 years)	*Math Shelf*	RCT	Active- other educational apps	Researcher developed	180	Positive towards app	.90	.64	14.0
Schacter & Jo (2016)	USA, School (Practitioner)	TD (162, 4-5 years)	*Math Shelf*	QED	BAU	Researcher developed	300	Positive towards app	1.22	.75	28.6
Schacter & Jo (2017)	USA, School (Practitioner)	TD (378, 4-5 years)	*Math Shelf*	RCT	Active – non-digital	Researcher developed	440	Positive towards app	.20	.79	12.7
Schaeffer et al. (2018)	USA, Home (Parent)	TD (195, 6-9 years)	*BedTime Math*	RCT	Active- other educational apps	Standardised- Woodcock-Johnson-III	NR	Positive towards app	NR	.82	66.8
Schenke et al. (2020)	USA, School (Researcher)	TD (99, 4-5 years)	*MeasureUp!*	RCT	Multiple	Researcher developed	240-360	Positive towards app	.65	.68	2.0
Spencer (2013)	UAE, School (NR)	TD (114, 4-5 years)	*Know Number Lite*	QED	BAU	Researcher developed	100	Positive towards app	.53	.54	28.8
Stacy et al. (2017)	USA, School (Practitioner)	TD (Not reported)	*IXL*	Qualitative	No control	Researcher developed	200	Mixed	NR	.75	NA
Stubbé et al. (2016)	Sudan, School (Practitioner)	TD (703, 7-9 years)	*E-Learning Sudan*	QED	BAU	Researcher developed	5,400	Positive towards app	NR	.68	23.3

Appendix 2: Summary of 50 included studies in the systematic review (Part 1)

Study	Country, Setting (Implementer)	Final Sample (n, age)	App(s)	Method	Controls	Primary Maths Outcome Measure	Time on Task (minutes)	Overall results	Primary Effect Size*	Quality Score	Attrition Rate %
Swicegood (2015)	USA, School (Practitioner)	TD (40, 6-7 years)	*Addimal Adventure; Splash Math 2nd Grade*	Mixed methods	No control	Researcher developed	4,800	Mixed	NR	.77	0
Tucker et al. (2016)	USA, School (Researcher)	TD (33, 7-8 years)	*6 apps*	Qualitative	No control	NA - engagement only	30-40	Positive towards app	NA	.70	0
Vanbecelaere et al. (2020)	Belgium, School (Practitioner)	TD (222, 6-7 years)	*Number Sense Game*	RCT	BAU	Researcher developed	300	Mixed	1.58	.71	33.9
Watts et al. (2016)	USA, School (Researcher)	see Litster et al. (2019)	*11 apps*	Mixed methods	No control	Researcher developed	see Litster et al. (2019)	Positive towards app	NA	.72	NA
Wu (2020)	USA, School (Researcher)	TD/LA (56, 4-5 years)	*MathAntics*	RCT	BAU	Researcher developed	180	Positive towards app	1.89	.68	11.1
Zander et al. (2016)	Germany, School (NR)	TD (37, 8-11 years)	*Rotate It!*	QED	Active – non-digital	Researcher developed	NR	Mixed	NR	.64	27.5
Zhang et al. (2020)	China, School (Practitioner)	TD (65, 8-9 years)	*Motion Math and Slice Fractions*	QED	BAU	Researcher developed	120	No difference	2.31	.64	14.5

**Within-subject effect sizes (Cohen's d) on primary mathematical outcome. Hedge's g correction applied for study samples equal to or less than 50 (reported in parenthesis).*

BAU= business-as-usual; IPE = implementation process evaluation; LA = low achievers; NA = not applicable; NR= not reported; QED = quasi-experimental design; RCT = randomised control trial; SEND = special educational needs and disabilities; TD = typically developing.

Appendix 3: Contributions of included studies (n =50) to themes identified in the gaps in the current evidence (Part 1)

Study	Near and far-transfer benefits of educational maths apps	Children underachieving in mathematics	Special educational needs and disabilities	Role of age and language in educational maths apps	Usage and immediate and sustained learning gains	Innovative methods for data collection	Cross-cultural comparisons	Educational maths apps in the home
Ahmad et al. (2014)			✓					
Berggren & Hedler (2014)								
Berkowitz et al. (2015)								✓
Broda et al. (2019)						✓		
Bullock et al. (2017)				✓				
Cary et al. (2020)		✓						
Cornu et al. (2019)	✓							
Ginsburg (2019)								
Griffith et al. (2019)								✓
Grimes et al. (2020)	✓							
Hasanah et al. (2017)						✓		
Hassler Hallstedt et al. (2018)	✓	✓			✓			
Hieftje et al. (2017)	✓	✓						
Hung et al. (2015)								
Judd & Klingberg (2021)						✓		
Kalmpourtzis (2014)								
Kosko & Ferdig (2016)								✓
Kromminga & Codding (2020)		✓						
Lee & Choi (2020)	✓							
Litster et al. (2019)								

Appendix 3: Contributions of included studies (n =50) to themes identified in the gaps in the current evidence (Part 1)

Study	Near and far-transfer benefits of educational maths apps	Children underachieving in mathematics	Special educational needs and disabilities	Role of age and language in educational maths apps	Usage and immediate and sustained learning gains	Innovative methods for data collection	Cross-cultural comparisons	Educational maths apps in the home
Mattoon et al. (2015)								
Miller (2018)								
Moyer-Packenham et al. (2016)								
Nunes et al. (2019)		✓					✓	
Outhwaite et al. (2017)		✓			✓		✓	
Outhwaite et al. (2018)	✓	✓					✓	
Outhwaite et al. (2019)							✓	
Outhwaite et al. (2020)				✓			✓	
Parks & Tortorelli (2020)								
Pecora (2015)			✓					
Pires et al. (2019)								
Pitchford (2015)	✓						✓	
Pitchford et al. (2018)			✓	✓		✓	✓	
Pitchford et al. (2019)							✓	
Ramani et al. (2019)	✓				✓			
Schacter & Jo (2016)								
Schacter & Jo (2017)	✓							
Schacter et al. (2016)								
Schaffer et al. (2018)					✓			✓
Schneke et al. (2020)								

Appendix 3: Contributions of included studies (n =50) to themes identified in the gaps in the current evidence (Part 1)

Study	Near and far-transfer benefits of educational maths apps	Children underachieving in mathematics	Special educational needs and disabilities	Role of age and language in educational maths apps	Usage and immediate and sustained learning gains	Innovative methods for data collection	Cross-cultural comparisons	Educational maths apps in the home
Spencer (2013)								
Stacy et al. (2017)								
Stubbé et al. (2016)								
Swicegood (2015)								
Tucker et al. (2016)								
Vanbecelaere et al. (2020)	✓				✓			
Watts et al. (2016)								
Wu (2020)		✓						
Zander et al. (2016)								
Zhang et al. (2020)	✓							

Appendix 4: Data collection form for each included app (Part 2)

Name of app	
Associated studies	
Type of app	Practice-based, Game-based, Constructive, Productive, Parent-based [delete as appropriate]
Free or paid app?	

Areas of maths covered within the app	**Dichotomously code (Not present [0] or present [1])**	**Description points (Not present [0] or present [1])**										
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Number representation and relationships												
Counting									▨	▨	▨	▨
Arithmetic												▨
Shape, patterns, and measurement												▨
Other (please state) (e.g., use DfE/ CCS/ EEF documents as reference if needed												

App design features	**Dichotomously code (Not present [0] or present [1])**	**Further details**
Feedback		What kind of feedback is provided (motivational, explanatory, or both)? How is the feedback delivered within the app (audio, visual, or both)?
Levelling		What kind of levelling is implemented within the app (programmatic static, programmatic dynamic, participatory free form, or a combination)?
Social interaction		What is the nature of the social interaction within the app? For apps with an in-app character, what kind of interactions are included (see Table 4 and/or others not listed)? For face to face/offline interactions, are supports provided for the adult to facilitate such interactions?
Task instruction		What opportunities are there for instructions to be repeated (i.e., how are they repeated)?
Meaningful learning and solving problems		What is the nature of the learning activities provided (see Table 4 for examples and/or others not listed)?
Other potential mechanisms of learning- within the app (please state)		Also refer to the associated paper for any features the authors feel may be important (in line with intervention component analysis principles; Sutcliffe et al., 2015). For example, quizzes in *onebillion* apps argued to align with principles of retrieval-based learning in associated papers.

Appendix 5: Summary of the Most Popular Commercially Available Maths Apps for 5-year-olds.

Name of App	**Type of App**	**Free or Paid**	**Areas of Mathematics**				**App Design Features**				
			Number	**Counting**	**Arithmetic**	**Shape, Patterns & Measurement**	**Feedback**	**Levelling**	**Social Interaction**	**Task Instruction**	**Meaningful Learning**
Kiddopia	Practice-based	Free (in-app subscription)	✓	✓	✓	✓		✓	✓		✓
ABC Mouse	Practice-based	Free (in-app subscription)		✓	✓	✓	✓	✓	✓	✓	✓
Meet the Numberblocks!	Constructive	Free	✓	✓				✓	✓		✓
Khan Academy Kids	Practice-based	Free	✓	✓	✓	✓	✓	✓	✓	✓	✓
Curious World: Learning Games	Practice-based	Free (in-app subscription)	✓	✓	✓	✓	✓	✓	✓		✓
Learning Games for Preschool	Practice-based	Free (in-app payments)	✓	✓			✓	✓	✓		✓
Toddler Games for Kids App Junior	Practice-based	Free (in-app payments)	✓	✓		✓	✓	✓	✓		✓
Hopster: Pre-school Games & TV	Practice-based	Free	✓	✓			✓	✓	✓		✓
Sago Mini School (Kids 2-5)	Practice-based	Free (in-app subscription)	✓	✓		✓		✓	✓		✓
BabyTV Video: Kids TV and Songs	Practice-based	Free	✓	✓		✓			✓		
Kids' Academy Educational Games for Kids 2 3	Practice-based	Free (in-app subscription)	✓	✓		✓	✓	✓	✓		✓
IntellectoKids Learning Games	Practice-based	Free (in-app subscription)	✓	✓	✓		✓	✓	✓	✓	✓
Baby Games for 2-5 Year Olds	Practice-based	Free (in-app subscription)				✓	✓	✓	✓		✓
Toddler Games for Kids 2+ Year	Practice-based	Free (in-app payments)	✓	✓		✓	✓	✓	✓		✓
TinyTap – Kids Learning Games	Practice-based	Free (in-app subscription)	✓	✓	✓	✓	✓	✓	✓		✓
Kids Academy: Pre-K-3 Learning	Practice-based	Free (in-app subscription)	✓	✓	✓	✓	✓	✓	✓		✓
Endless Numbers	Practice-based	Free	✓	✓	✓		✓	✓	✓		✓
Interactive Time Telling	Practice-based	Free				✓	✓	✓	✓	✓	✓

Appendix 6: Educational apps identified as 'maths' on the search terms that did not explicitly include any maths content.

The following six apps did not contain any mathematical content:

- Toddler Games: 3,4,5 Year Olds
- Dinosaur Park: Games for Kids
- Toddler Games for 3+ Year Old
- Kids Fun Preschool Maths Games
- Dinosaur Airport: Airplane fly
- Games for 3, 4, 5, 6, 7 Year Olds

References

Included Studies (n = 50)

Ahmad, W. F. W., Muddin, H. N. B. I., & Shafie, A. (2014). Number skills mobile application for down syndrome children. In *2014 International Conference on Computer and Information Sciences (ICCOINS)* (p. 1-6).

Berggren, J., & Hedler, C. (2014). CamQuest: Design and evaluation of a tablet application for educational use in preschools. In *Proceedings of the 2014 conference on Interaction design and children* (p. 185-188).

Berkowitz, T., Schaeffer, M. W., Maloney, E. A., Peterson, L., Gregor, C., Levine, S. C., & Beilock, S. L. (2015). Math at home adds up to achievement in school. *Science*, *350*(6257), 196-198.

Broda, M., Tucker, S., Ekholm, E., Johnson, T. N., & Liang, Q. (2019). Small fingers, big data: Preschoolers' subitizing speed and accuracy during interactions with multitouch technology. *The Journal of Educational Research*, *112*(2), 211-222.

Bullock, E. P., Shumway, J. F., Watts, C. M., & Moyer-Packenham, P. S. (2017). Affordance access matters: Preschool children's learning progressions while interacting with touch-screen mathematics apps. *Technology, Knowledge and Learning*, *22*(3), 485-511.

Cary, M. S., Kennedy, P. C., Shanley, L., & Clarke, B. (2020). Learning Gains From the KinderTEK® iPad Math Program: Does Timing of a Preventative Intervention Matter?. *Journal of Special Education Technology*, *36*(4), 321-335.

Cornu, V., Schiltz, C., Pazouki, T., & Martin, R. (2019). Training early visuo-spatial abilities: A controlled classroom-based intervention study. *Applied Developmental Science*, *23*(1), 1-21. *Knowledge and Learning*, *22*(3), 485-511.

Ginsburg, H. P., Wu, R., & Diamond, J. S. (2019). MathemAntics: a model for computer-based mathematics education for young children/MathemAntics: un modelo de enseñanza de matemáticas asistida por ordenador para

Griffith, S. F., Hanson, K. G., Rolon-Arroyo, B., & Arnold, D. H. (2019). Promoting early achievement in low-income preschoolers in the United States with educational apps. *Journal of Children and Media*, *13*(3), 328-344.

Grimes, K. R., Park, S., McClelland, A., Park, J., Lee, Y. R., Nozari, M., ... & Zaparolli, B. (2020). Effectiveness of a Numeracy Intelligent Tutoring System in Kindergarten: A Conceptual Replication. *Open Science Framework Pre-print.* Available from: https://osf.io/n2x7b/download.

Hasanah, N., Hayashi, Y., & Hirashima, T. (2017). An analysis of learner outputs in problem posing as sentence-integration in arithmetic word problems. *Research and practice in technology enhanced learning*, *12*(1), 1-16.

Hassler Hallstedt, M., Klingberg, T., & Ghaderi, A. (2018). Short and long-term effects of a mathematics tablet intervention for low performing second graders. *Journal of Educational Psychology*, *110*(8), 1127.

Hieftje, K., Pendergrass, T., Kyriakides, T. C., Gilliam, W., & Fiellin, L. (2017). An evaluation of an educational video game on mathematics achievement in first grade students. *Technologies*, *5*(2), 30.

Hung, C. Y., Sun, J. C. Y., & Yu, P. T. (2015). The benefits of a challenge: student motivation and flow experience in tablet-PC-game-based learning. *Interactive Learning Environments*, *23*(2), 172-190.

Judd, N., & Klingberg, T. (2021). Training spatial cognition enhances mathematical learning in a randomized study of 17,000 children. *Nature Human Behaviour*, 1-7.

Kalmpourtzis, G. (2014). Teaching of Spatial Thinking in Early Childhood Through Game-Based Learning: The use of the Ipad. In *InECGBL20 14–8th European Conference on Games Based Learning: ECGBL 2014* (p. 231).

Kosko, K., & Ferdig, R. (2016). Effects of a tablet-based mathematics application for pre-school children. *Journal of Computers in Mathematics and Science Teaching*, *35*(1), 61-79.

Kromminga, K. R., & Codding, R. S. (2020). A comparison of 1: 1 flashcards and a tablet app on student mathematics proficiency. *Journal of Behavioral Education*, 1-26.

Lee, H. K., & Choi, A. (2020). Enhancing early numeracy skills with a tablet-based math game intervention: a study in Tanzania. *Educational Technology Research and Development*, *68*(6), 3567-3585.

Litster, K., Moyer-Packenham, P. S., & Reeder, R. (2019). Base-10 blocks: A study of iPad virtual manipulative affordances across primary-grade levels. *Mathematics Education Research Journal*, *31*(3), 349-365.

Mattoon, C., Bates, A., Shifflet, R., Latham, N., & Ennis, S. (2015). Examining Computational Skills in Prekindergarteners: The Effects of Traditional and Digital Manipulatives in a Prekindergarten Classroom. *Early Childhood Research & Practice*, *17*(1), n1.

Miller, T. (2018). Developing numeracy skills using interactive technology in a play-based learning environment. *International Journal of STEM Education*, *5*(1), 1-11.

Moyer-Packenham, P. S., Bullock, E. K., Shumway, J. F., Tucker, S. I., Watts, C. M., Westenskow, A., ... & Jordan, K. (2016). The role of affordances in children's learning performance and efficiency when using virtual manipulative mathematics touch-screen apps. *Mathematics Education Research Journal*, *28*(1), 79-105.

Nunes, T., Malmberg, L. E., Evans, D., Sanders-Ellis, D., Baker, S., Barros, R., ... & Evangelou, M. (2019). *onebillion: Evaluation Report.* Education Endowment Foundation (EEF): London, UK.

Outhwaite, L. A., Gulliford, A., & Pitchford, N. J. (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.

Outhwaite, L. A., Faulder, M., Gulliford, A., & Pitchford, N. J. (2018). Raising early achievement in math with interactive apps: A randomized control trial. *Journal of Educational Psychology*, *111*(2), 284-298.

Outhwaite, L. A., Gulliford, A., & Pitchford, N. J. (2019). A new methodological approach for evaluating the impact of educational intervention implementation on learning outcomes. *International Journal of Research & Method in Education*, *43*(3), 225-242.

Outhwaite, L. A., Gulliford, A., & Pitchford, N. J. (2020). Language counts when learning mathematics with interactive apps. *British Journal of Educational Technology*, *51*(6), 2326-2339.

Parks, A.N. and Tortorelli, L. (2020). Impact of a district-wide one-to-one technology initiative on kindergartners' engagement and learning outcomes. *Journal of Research in Childhood Education*, doi: 10.1080/02568543.2020.1809578.

Pecora, J. (2015). *What is the effect of specific technology apps on student achievement, focus, and attention in a Kindergarten Special Education Mathematics classroom? (Masters Dissertation).* New Jersey: Caldwell University.

Pires, A. C., González Perilli, F., Bakała, E., Fleisher, B., Sansone, G., & Marichal, S. (2019). Building blocks of mathematical learning: Virtual and tangible manipulatives lead to different strategies in number composition. *Frontiers in Education*, *4*(81), doi:10.3389/feduc.2019.00081.

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.

Pitchford, N.J., Kamchedzera, E., Hubber, P.J., & Chigeda, A.L. (2018). Interactive apps promote learning of basic mathematics in children with Special Educational Needs and Disabilities. *Frontiers in Psychology, 9*(262), doi: 10.3389/fpsyg.2018.00262.

Pitchford, N. J., Chigeda, A., & Hubber, P. J. (2019). Interactive apps prevent gender discrepancies in early-grade mathematics in a low-income country in sub-Sahara Africa. *Developmental Science*, *22*(5), e12864.

Ramani, G. B., Daubert, E. N., Lin, G. C., Kamarsu, S., Wodzinski, A., & Jaeggi, S. M. (2020). Racing dragons and remembering aliens: Benefits of playing number and working memory games on kindergartners' numerical knowledge. *Developmental Science*, *23*(4), e12908.

Schacter, J., Shih, J., Allen, C. M., DeVaul, L., Adkins, A. B., Ito, T., & Jo, B. (2016). Math shelf: A randomized trial of a prekindergarten tablet number sense curriculum. *Early Education and Development*, *27*(1), 74-88.

Schacter, J., & Jo, B. (2016). Improving low-income preschoolers mathematics achievement with Math Shelf, a preschool tablet computer curriculum. *Computers in Human Behavior*, *55*, 223-229.

Schacter, J., & Jo, B. (2017). Improving preschoolers' mathematics achievement with tablets: A randomized controlled trial. *Mathematics Education Research Journal*, *29*(3), 313-327.

Schaeffer, M. W., Rozek, C. S., Berkowitz, T., Levine, S. C., & Beilock, S. L. (2018). Disassociating the relation between parents' math anxiety and children's math achievement: Long-term effects of a math app intervention. *Journal of Experimental Psychology: General*, *147*(12), 1782-1790.

Schenke, K., Redman, E. J., Chung, G. K., Chang, S. M., Feng, T., Parks, C. B., & Roberts, J. D. (2020). Does "Measure Up!" measure up? Evaluation of an iPad app to teach preschoolers measurement concepts. *Computers & Education*, *146*, 103749.

Spencer, P. (2013). iPads: Improving numeracy learning in the early years. In *Proceedings of the 36th Annual Conference of the Mathematics Education Research Group of Australasia (MERGA)* (p.610-617).

Stacy, S. T., Cartwright, M., Arwood, Z., Canfield, J. P., &

Kloos, H. (2017). Addressing the math-practice gap in elementary school: Are tablets a feasible tool for informal math practice?. *Frontiers in Psychology*, *8*,(179), doi:10.3389/fpsyg.2017.00179.

Stubbé, H., Badri, A., Telford, R., van der Hulst, A., & van Joolingen, W. (2016). E-Learning Sudan, Formal Learning for Out-of-School Children. *Electronic Journal of e-Learning*, *14*(2), 136-149.

Swicegood, G.P. (2015). *An investigation of the impact of iPad usage on elementary mathematical skills and attitudes (PhD Thesis).* Missoula: University of Montana.

Tucker, S. I., Moyer-Packenham, P. S., Westenskow, A., & Jordan, K. E. (2016). The complexity of the affordance–ability relationship when second-grade children interact with mathematics virtual manipulative apps. *Technology, Knowledge and Learning*, *21*(3), 341-360.

Vanbecelaere, S., Van den Berghe, K., Cornillie, F., Sasanguie, D., Reynvoet, B., & Depaepe, F. (2020). The effects of two digital educational games on cognitive and non-cognitive math and reading outcomes. *Computers & Education*, *143*, 103680.

Watts, C. M., Moyer-Packenham, P. S., Tucker, S. I., Bullock, E. P., Shumway, J. F., Westenskow, A., ... & Jordan, K. (2016). An examination of children's learning progression shifts while using touch screen virtual manipulative mathematics apps. *Computers in Human Behavior*, *64*, 814-828.

Wu, R. (2020). The effect of the math application *MathemAntics* on preschoolers' math performance (*PhD Thesis).* New York: Columbia University.

Zander, S., Wetzel, S., & Bertel, S. (2016). Rotate it!–Effects of touch-based gestures on elementary school students' solving of mental rotation tasks. *Computers & Education*, *103*, 158-169.

Zhang, L., Shang, J., Pelton, T., & Pelton, L. F. (2020). Supporting primary students' learning of fraction conceptual knowledge through digital games. *Journal of Computer Assisted Learning*, *36*(4), 540-548.

Additional References

American Academy of Pediatrics. (2016) *American Academy of Pediatrics Announces New Recommendations for Children's Media Use.* Available from: https://www.aap.org/en-us/about-the-aap/aap-press-room/Pages/American-Academy-of-Pediatrics-Announces-New-Recommendations-for-Childrens-Media-Use.aspx.

Apple. (2014). *iPad in Education Results*. 1–19.

Barnett, A. G., Van Der Pols, J. C., & Dobson, A. J. (2004). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215-220.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101.

Butterworth, B. (2005). The development of arithmetical abilities. *Journal of Child Psychology and Psychiatry*, *46*(1), 3-18.

Callaghan, M. N., & Reich, S. M. (2018). Are educational preschool apps designed to teach? An analysis of the app market. *Learning, Media and Technology*, *43*(3), 280-293.

Callaghan, M. N., & Reich, S. M. (2021). Mobile app features that scaffold pre-school learning: Verbal feedback and leveling designs. *British Journal of Educational Technology*, *52*(2), 785-806.

Chaudron, S. (2015). *Young Children (0-8) and Digital Technology a Qualitative Exploratory Study across Seven Countries.* Available from: http://www.lse.ac.uk/media@lse/research/ToddlersAndTablets/RelevantPublications/Young-Children-(0-8)-and-Digital-Technology.pdf.

Cheung, A. C., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review*, *9*, 88-113.

Clark, A., Henderson, P., Gifford, S., Hodgen, J., Barclay, N., Foster, C., ... & Simms, V. (2020). *Improving Mathematics in the Early Years and Key Stage 1.* Available from: https://educationendowmentfoundation.org.uk/public/files/Publications/Maths/EEF_Maths_EY_KS1_Guidance_Report.pdf.

Clements, D. H. & Sarama, J. (2009). *Learning and Teaching Early Math.* London: Routledge.

Common Core State Standards Initiative. (n.d.). *Common Core State Standards for Mathematics.* Available from: http://www.corestandards.org/wp-content/uploads/Math_Standards1.pdf

Department for Education. (2013). *Mathematics Programmes of Study: Key Stages 1 and 2.* Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/335158/PRIMARY_national_curriculum_-_Mathematics_220714.pdf.

Department for Education. (2017). *Bold Beginnings.* Available from: https://www.gov.uk/government/publications/reception-curriculum-in-good-and-outstanding-primary-schools-bold-beginnings.

Department for Education. (2020). *SEND code of practice: 0 to 25 years.* Available from: https://www.gov.uk/government/publications/send-code-of-practice-0-to-25.

Department for Education. (2021). *Statutory Framework for the Early Years Foundation Stage.* Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/974907/EYFS_framework_-_March_2021.pdf.

Department for Work and Pensions (2019). *Households below average income: 1994/95 to 2017/18.* Available from: https://www.gov.uk/government/statistics/households-below-average-income-199495-to-201718

Education Technology. (2021). *UK edtech market expected to reach £3.4bn this year.* Available from: https://edtechnology.co.uk/business/uk-edtech-market-expected-reach-3-4bn-this-year/.

Falloon, G. (2013). Young students using iPads: App design and content influences on their learning pathways. *Computers & Education*, *68*, 505-521.

Gilmore, C., Göbel, S. M., & Inglis, M. (2018). *An Introduction to Mathematical Cognition*. London: Routledge.

Green, C. S., Bavelier, D., Kramer, A. F., Vinogradov, S., Ansorge, U., Ball, K. K., ... & Witt, C. M. (2019). Improving methodological standards in behavioral interventions for cognitive enhancement. *Journal of Cognitive Enhancement*, *3*(1), 2-29.

Griffith, S. F., Hagan, M. B., Heymann, P., Heflin, B. H., & Bagner, D. M. (2020). Apps as learning tools: a systematic review. *Pediatrics*, *145*(1).

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.

Herodotou, C. (2021). MAD learn: An evidence-based affordance framework for assessing learning apps. *Immersive Learning Research Network.*

Hirsh-Pasek, K., Zosh, J. M., Golinkoff, R. M., Gray, J. H., Robb, M. B., & Kaufman, J. (2015). Putting Education in "Educational" Apps: Lessons from the Science of Learning. *Psychological Science*, *16*(1), 3-34.

Hsin, C. T., & Wu, H. K. (2011). Using scaffolding strategies to promote young children's scientific understandings of floating and sinking. *Journal of Science Education and Technology*, *20*(5), 656-666.

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

Jordan, N.C. & Levine, S.C. (2009). Socioeconomic variation, number competence, and mathematics learning difficulties in young children. *Developmental Disabilities Research Reviews*. *15*(1), 60-68.

Kay, R., & Kwak, J. Y. (2018). Comparing types of mathematics apps used in primary school classrooms: an exploratory analysis. *Journal of Computers in Education*, *5*(3), 349-371.

Kim, J., Gilbert, J., Yu, Q., & Gale, C. (2021). Measures Matter: A Meta-Analysis of the Effects of Educational Apps on Preschool to Grade 3 Children's Literacy and Math Skills. *AERA Open*, *7*, 23328584211004183.

Kmet, L. M., Cook, L. S., & Lee, R. C. (2004). *Standard quality assessment criteria for evaluating primary research papers from a variety of fields.* Available from: https://era.library.ualberta.ca/items/48b9b989-c221-4df6-9e35-af782082280e.

Kolak, J., Norgate, S. H., Monaghan, P., & Taylor, G. (2021). Developing evaluation tools for assessing the educational potential of apps for preschool children in the UK. *Journal of Children and Media*, *15*(3), 410-430.

Kucirkova, N. (2018). Personalised Learning with Digital Technologies at Home and School: Where is Children's Agency?. In *Mobile Technologies in Children's Language and Literacy*. Bingley: Emerald Publishing Limited.

Lin, L. (2018). Bias caused by sampling error in meta-analysis with small sample sizes. *PloS one*, *13*(9), e0204056.

London & Partners. (2015). *Edtech: London Capital for Learning Technology*. Available from: http://www.ednfoundation.org/wp-content/uploads/EdtechUK_LP_report.pdf.

Magliaro, S. G., Lockee, B. B., & Burton, J. K. (2005). Direct instruction revisited: A key model for instructional technology. *Educational Technology Research and Development*, *53*(4), 41-55.

Maloney, E. A., & Beilock, S. L. (2012). Math anxiety: Who has it, why it develops, and how to guard against it. *Trends in Cognitive Sciences*, *16*(8), 404-406.

Marsh, J., Murris, K., Ng'ambi, D., Parry, R., Scott, F., Thomsen, B. S., ... & Woodgate, A. (2020). *Children, Technology and Play.* Billund: The LEGO Foundation.

Meyer, M., Zosh, J. M., McLaren, C., Robb, M., McCaffery, H., Golinkoff, R. M., ... & Radesky, J. (2021). How educational are "educational" apps for young children? App store content analysis using the Four Pillars of Learning framework. *Journal of Children and Media*, 1-23.

Organisation for Economic Co-operation and Development. (2020). *Early Learning and Child Well-being: A Study of Five-year-Olds in England, Estonia, and the United States.* Available from: https://www.oecd.org/education/school/early-learning-and-child-well-being-3990407f-en.htm.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Bmj*, *372*(71), doi: 10.1136/bmj.n71.

Papadakis, S., Kalogiannakis, M., & Zaranis, N. (2017). Designing and creating an educational app rubric for preschool teachers. *Education and Information Technologies*, *22*(6), 3147-3165.

Pawson, R., & Tilley, N. (1997). Realistic evaluation. In S. Matthieson (Ed.), *Encyclopedia of Evaluation* (p. 359-367). London: Sage.

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin*, *135*(6), 943-973.

Sella, F., Onnivello, S., Lunardon, M., Lanfranchi, S., & Zorzi, M. (2021). Training basic numerical skills in children with Down syndrome using the computerized game "The Number Race". *Scientific Reports*, *11*(1), 1-14.

Sella, F., Tressoldi, P., Lucangeli, D., & Zorzi, M. (2016). Training numerical skills with the adaptive videogame "The Number Race": A randomized controlled trial on preschoolers. *Trends in Neuroscience and Education*, *5*(1), 20-29.

Simms, V., McKeaveney, C., Sloan, S. & Gilmore, C. (2019). *Interventions to Improve Mathematical Achievement in Primary School-Aged Children*. Available from: https://www.ulster.ac.uk/__data/assets/pdf_file/0003/417864/web-00553-05_UU_A4_Report_v7.pdf.

Stigler, S. M. (1997). Regression towards the mean, historically considered. *Statistical Methods in Medical Research*, *6*(2), 103-114.

Sutcliffe, K., Thomas, J., Stokes, G., Hinds, K., & Bangpan, M. (2015). Intervention Component Analysis (ICA): a pragmatic approach for identifying the critical features of complex interventions. *Systematic Reviews*, *4*(1), 140.

Thomas, J., O'Mara-Eves, A., & Brunton, G. (2014). Using qualitative comparative analysis (QCA) in systematic reviews of complex interventions: a worked example. *Systematic Reviews*, *3*(1), 1-14.

Toth, K., Sammons, P., Sylva, K., Melhuish, E., Siraj, I., & Taggart, B. (2020). Home learning environment across time: the role of early years HLE and background in predicting HLE at later ages. *School Effectiveness and School Improvement*, *31*(1), 7-30.

Tucker, S. I. (2015). *An Exploratory Study of Attributes, Affordances, Abilities, and Distance in Children's Use of Mathematics Virtual Manipulative iPad Apps (Doctoral dissertation).* Utah: Utah State University.

United Nations. (2021). *The Sustainable Development Goals Report.* Available from: https://unstats.un.org/sdgs/report/2021/The-Sustainable-Development-Goals-Report-2021.pdf

Vanbecelaere, S., Cornillie, F., Sasanguie, D., Reynvoet, B., & Depaepe, F. (2021). The effectiveness of an adaptive digital educational game for the training of early numerical abilities in terms of cognitive, noncognitive and efficiency outcomes. *British Journal of Educational Technology*, *52*(1), 112-124.

Vandewaetere, M., & Clarebout, G. (2014). Advanced technologies for personalized learning, instruction, and performance. In *Handbook of research on educational communications and technology* (p. 425-437). New York: Springer.

Yudkin, P. L., & Stratton, I. M. (1996). How to deal with regression to the mean in intervention studies. *Lancet*, *347*, 241-243.

Accessed apps

Accessed Maths Apps

Aleesha Kondys. *Friends of Ten* (Version 1.2) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/friends-of-ten/id488573871

Aleesha Kondys. *Teaching Number Lines* (Version 2.0) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/teaching-number-lines/id492603378

Bedtime Math Foundation. *Bedtime Math* (Version 5.2) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/bedtime-math/id637910701

Clockwork Fox inc. *Zorbit's Math Adventure* (1.11.1) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/KinderTEK-offline/id986096168

Cognition Matters. *Vektor* (Version 1.8.11) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/vektor/id1108806894

Herbert Ginsburg. *MathemAntics* (Version 1.1) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/mathemantics/id1509464562

Higher Ground Education Inc. *Montessorium: Intro to Math* (Version 3.1.2) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/montessorium-intro-to-math/id381064973

Image & Form International AB. *Fingu* (Version 2.7) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/fingu/id449815506?l=en

IXL Learning. *IXL- Math, English, & More* (Version 7.1.0) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/ixl-math-english-more/id693689912

John Schacter. *Math Shelf: Early Math Mastery* (Version 4.0) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/math-shelf-early-math-mastery/id940278648

Kittipong Mungnirun. *Know Number Lite* (Version 2.1) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/know-number-lite/id488062813

L'Escapadou. *Montessori Numbers for Kids* (Version 3.1.1) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/montessori-numbers-for-kids/id547209737

Matthew Thomas. *100s Board* (Version 1.7) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/100s-board/id372574961

MontessoriTech. *Montessori Bead Skip Counting* (Version 1.5) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/montessori-bead-skip-counting/id523627488

Native Brain, Inc. *Native Numbers* (Version 1.53) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/native-numbers/id570231808

onebillion. *Maths, age 3-5* (Version 1.11) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/maths-age-3-5/id471864743

onebillion. *Maths, age 4-6* (Version 3.7) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/maths-age-4-6/id509457434

Quizlet Inc. *Quizlet Plus* (Version 6.9.1) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/quizlet-learn-with-flashcards/id546473125

Rantek Inc. *Pink Tower- Montessori Math* (Version 2.3) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/pink-tower-montessori-math/id407599645

StudyPad, Inc. *Splash Math: Year 2 Maths learning games* (Version 5.6.1) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/2nd-grade-math-learning-games/id463471155

Teachley. *Addimal Adventure* (Version 2.10) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/teachley-addimal-adventure/id661286973

Ululab. *Slice Fractions* (Version 1.04.09) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/slice-fractions/id794730213

University of Oregon Digital Press. *KinderTEK Offline* (Version 2.5.3) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/KinderTEK-offline/id986096168

Accessed Apps (Top 25 most popular commercial educational maths apps)

Age of Learning, Inc. *ABC Mouse* (Version 8.33.0). [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/abcmouse-com/id586328581

Amaya Soft MChJ. *Baby Games for 2-5 Year Olds* (Version 1.6.0) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/baby-games-for-2-5-year-olds/id1539389424

Baby Network Limited. *BabyTV Video: Kids TV & Songs* (Version 6.1.2) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/babytv-video-kids-tv-songs/id428267291

Blue Zoo. *Meet the NumberBlocks!* (Version 1.1.0) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/meet-the-numberblocks/id1445555400

Funny Food: Kids Learning Games. *Kids' Academy Educational Games for Kids 2 3* (Version 2.6.0) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/educational-kids-maths-games/id1447377766

GiggleUp Kids Apps And Educational Games Pty Ltd. *Interactive Telling Time* (Version 3.7) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/interactive-telling-time/id482452233

Hopster. *Hopster: Pre-school Games & TV* (Version 3.57.30) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/hopster-pre-school-games-tv/id689165391

IntellectoKids Ltd. *IntellectoKids Learning Games* (Version 4.12.2) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/intellectokids-learning-games/id1184237977

Khan Academy. *Khan Academy Kids* (Version 4.2) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/khan-academy-kids/id1378467217

Kids Academy Co apps: Preschool & Kindergarten Learning Kids Games, Educational Books, Free Songs. *Kids Academy: Pre-K-3 Learning* (Version 8.2.1) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/kids-academy-pre-k-3-learning/id543851593

Kidsy Limited. *Curious World: Games for Kids* (Version 2.37) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/curious-world-games-for-kids/id994556645

Originator Inc. *Endless Numbers* (Version 3.1) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/endless-numbers/id804360921

Paper Boat Apps. *Kiddopia* (Version 4.4.1) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/kiddopia-abc-toddler-games/id1223397201

Sago Mini. *Sago Mini School (Kids 2-5)* (Version 2.4) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/sago-mini-school-kids-2-5/id1483068197

Skidos Learning. *Games for 3,4,5,6,7 Year Olds* (Version 40) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/games-for-3-4-5-6-7-year-olds/id1407833333

Skidos Learning. *Kids Fun Preschool Maths Games* (Version 3.6) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/kids-fun-preschool-maths-games/id1497549298

Skidos Ltd. *Toddler Games for 3+ Year Old* (Version 5.2) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/toddler-games-for-3-year-old/id1506886061

Skidos Ltd. *Toddler Games: 3,4,5 Year Olds* (Version 5.1) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/toddler-games-3-4-5-year-olds/id1483744837

TinyTap Ltd. *TinyTap – Kids Learning Games* (Version 4.1.2) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/tinytap-kids-learning-games/id493868874

TOYA TAP: PRESCHOOL AND KINDERGARTEN FREE LEARNING PUZZLES GAMES FOR KIDS TODDLERS LTD. *Learning games for preschool* (Version 1.19.1) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/learning-games-for-preschool/id1176157709

TOYA TAP: PRESCHOOL AND KINDERGARTEN FREE LEARNING PUZZLES GAMES FOR KIDS TODDLERS LTD. *Toddler Games for Kids 2+ Year* (Version 5.1) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/toddler-games-for-kids-2-year/id1171350534

TOYA TAP: PRESCHOOL AND KINDERGARTEN FREE LEARNING PUZZLES GAMES FOR KIDS TODDLERS LTD. *Toddler Games for Kids App Junior* (Version 1.12) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/toddler-games-for-kids-apps-jr/id1176162006

Yateland Limited. *Dinosaur Airport: Airplane fly* (Version 1.0.4) [Mobile app]. Apple App Store. https://apps.apple.com/gb/app/dinosaur-airport-airplane-fly/id1448993311

Yateland Limited. *Dinosaur Park: Games for Kids* (Version 1.0.3) [Mobile app]. Apple App Store. https://apps.apple.com/us/app/dinosaur-park-games-for-kids/id1071759290