Routledge
Taylor & Francis Group

Research Papers in Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rred20

Is technology always helpful?: A critical review of the impact on learning outcomes of education technology in supporting formative assessment in schools

Beng Huat See, Stephen Gorard, Binwei Lu, Lan Dong & Nadia Siddiqui

To cite this article: Beng Huat See, Stephen Gorard, Binwei Lu, Lan Dong & Nadia Siddiqui (2022) Is technology always helpful?: A critical review of the impact on learning outcomes of education technology in supporting formative assessment in schools, Research Papers in Education, 37:6, 1064-1096, DOI: 10.1080/02671522.2021.1907778

To link to this article: https://doi.org/10.1080/02671522.2021.1907778

Published online: 29 Mar 2021.

RESEARCH PAPERS IN EDUCATION
2022, VOL. 37, NO. 6, 1064–1096
https://doi.org/10.1080/02671522.2021.1907778

Routledge
Taylor & Francis Group

Is technology always helpful?: A critical review of the impact on learning outcomes of education technology in supporting formative assessment in schools

Beng Huat See, Stephen Gorard, Binwei Lu, Lan Dong and Nadia Siddiqui

School of Education, Durham University, Durham, UK

ABSTRACT

While education technology has been widely used in classrooms, and considerable investments have been made to support its use in the UK, the evidence base for many such rapidly changing technologies is weak, and their efficacy is unclear. The aim of this paper is to systematically review and synthesise empirical research on the use of technology in formative assessment, to identify approaches that are effective in improving pupils' learning outcomes. The review involved a search of 11 major databases, and included 55 eligible studies. The results suggest promising evidence that digitally delivered formative assessment could facilitate the learning of maths and reading for young children, but there is no good evidence that it is effective for other subjects, or for older children, or that it is any more effective than formative assessment without technology. The review found no good evidence that learner response systems work in enhancing children's academic attainment, and there is no evidence supporting the effectiveness of such technologies that embed gaming features. Much research in this area is of poor quality. More rigorous studies using causal designs are thus urgently needed. Meantime, there should be no rush to use technology on the basis of improving attainment.

ARTICLE HISTORY
Received 27 December 2020
Accepted 21 March 2021

KEYWORDS
Education technology; formative assessment; systematic review

Background

There is increasing interest in the use of technology in education. In April 2019, the UK government announced a £10 m investment in education technology for England. This new strategy was intended to tackle common challenges in using educational technology, equip teachers with the necessary skills and resources, and to support teachers by reducing workload (Department for Education, [DfE], 2019).

There had also been several schemes in the past offering disadvantaged children home computers. The Computers for Pupils scheme (Department for Education and Skills [DfES], 2006), for instance, distributed funding of more than £60 million for pupils aged 11–16 in the most deprived areas of England, mainly focusing on the use of computers for homework and providing course materials online (Lynch et al. 2010). Evaluation of its impact was essentially based on surveys of teachers', parents' and pupils' perceptions. Impact on attainment could not be ascertained because only 81 pupils who could also be

CONTACT Beng Huat See b.h.see@durham.ac.uk School of Education, Durham University, UK

matched to the National Pupil Database said they had received a computer. And, in 2008, the English government committed £300 million to the Home Access programme, which funded computers and internet access for pupils aged 8–19 in disadvantaged families (Becta 2009). In schools a wide range of technology, gadgets, and commercial software programmes have become available to teachers to support teaching and learning. And, although this is beyond the scope of this paper, the appeal of digital educational platforms increased in 2020 due to the Covid-19 lockdown.

While there have been ongoing efforts by governments to enhance the use of digital technology (DT) or education technology (EdTech) in schools, there is no clear evidence yet that the use of technology alone can lead to improvements in learning outcomes (e.g. Luckin et al. 2012; Gorard, See, and Morris 2016; US Department of Education 2014). OECD (2015) found no link between the amount of technology use in class and students' performance on PISA (Programme for International Assessment). Higgins, Xiao, and Katsipataki (2012) also noted that the link between educational technology and attainment may not be causal or even linear. This has prompted high-profile concerns over how digital technologies are being used in schools. OECD's education director Andreas Schleicher observed in 2015 that 'even where computers are used in the classroom, their impact on student performance is mixed, at best' (Schleicher 2015).

Embedding appropriate technology in the curriculum can be challenging. Some suggested reasons for this include a lack of leadership and management (Office for Standards in Education [Ofsted], 2011), and insufficient coordination and lack of support, alongside resourcing issues and constraints of curriculum and assessment requirements. A report by a software technology company, *Driving Digital Strategy in Schools*, pointed to the fragmented guidance given to schools, and a lack of a coherent evidence base (Instructure 2017). While some countries have consulted schools to identify the challenges and opportunities they faced in the effective use of EdTech (e.g. the Irish school review by Cosgrove et al. 2014), no similar widespread consultative work has been conducted for England.

It is generally recognised that EdTech can facilitate the learning of maths and science (if used effectively) and some aspects of literacy, but this impact may be no better than other interventions not involving technologies (The Scottish Government 2015; Gorard, See, and Morris 2016; Lewin et al. 2019). The aim of this paper is to review the international evidence on the impact on young peoples' learning outcomes of the use of EdTech in supporting formative assessment. The paper is based on a study funded by Nesta and the Department of Education to evaluate the use of EdTech.

Previous reviews

There have been several systematic reviews and meta-analyses looking at a range of educational technology (EdTech) to improve learning (e.g. Higgins, Xiao, and Katsipataki 2012), including game-based learning and use of videos (e.g. Byun and Joung 2018; Tokac, Novak, and Thompson 2019), with some focusing on specific curriculum subjects, such as maths (Cheung and Slavin 2013), reading (Cheung and Slavin 2012) and English as a foreign language (Chiu, Kao, and Reynolds 2012). Despite it being widely used, there have been few reviews or meta-analyses focused on online formative assessment or the learner response system for school children (Lewin et al.

2019). Most that exist are based on higher education. We found only one meta-analysis on the use of audience response system (or clickers) in a language classroom (Castillo-Manzano et al. 2016). Another focused on the use of automated feedback in a programming class (Keuning, Jeuging & Heeren 2018). Shute and Rahimi (2017) reviewed computer-based assessment for learning (CBAfL) across a range of content areas, including maths biology and programming, reporting different effect sizes for each area.

Problems with previous reviews

Previous reviews of the use of EdTech have been largely syntheses of meta-analyses, and most of the underlying reviews are based on a summary of effect sizes from individual studies. These tended to suggest evidence of the effectiveness of EdTech. However, effect sizes are driven partly by sample size and research design. Weaker studies with small samples, using non-randomised controls, or instruments that measure outcomes closely related to the content of the intervention or based on teachers'/pupils' self-reports of outcomes, tend to show suspiciously large apparent effect sizes (Slavin and Smith 2009; Slavin and Madden 2011; Van der Kleij, Feskens and Eggen 2015). Therefore, averaging effect sizes from reviews which pooled the effects from different individual studies of disparate quality has the potential for errors in the estimates to be propagated. As a result, some meta-analyses have reported unrealistically large effect sizes. Hattie's meta-analysis (Hattie 2018), for example, reported an effect size of +0.42 for the use of technology in writing. Sung, Chang, and Liu (2016) reported effect sizes ranging from g = +0.45 to g = +0.78 for mobile learning in maths, science, literacy and language with effects varying depending on age group and settings. However, almost half of the studies in the review did not have a control group, but were treated in the same way as the remaining studies. This is not an appropriate summary of evidence.

Another issue with previous reviews is that they have aggregated studies for a wide age range of learners, from early years to undergraduates and adults. Effect sizes of studies that include adults tend to report larger effects. The meta-analysis by Kulik and Fletcher (2016) showed an effect of g = +0.44 for school learners and g = +0.78 for adults. Chen, Tseng, and Hsiao (2018) reported effect sizes of +1.87 for the use of adventure games, but two of the three studies were with adults. Zheng's (2018) study of mobile devices and inquiry-based learning suggests much bigger effects for university students (e.g. medical science 1.82, natural science 0.93, social science 0.57).

Previous meta-analyses of reviews also did not take into account the quality of individual studies in the reviews they meta-analysed. Lewin et al. (2019), for example rated Chen, Tseng & Hsiao's (2018) review as moderate with regards to certainty of effects based on the Maryland Scientific Methods Scale. One study in their review (Saffarian and Gorjian 2012) suggested an effect of d = +2.3. This is what Slavin equates to finding a 10-foot giant. The study could not be located by us, despite an interlibrary loan request, so we were unable to assess the quality of this study, but another study (Vahdat and Behbahani 2013) was described as having an effect of d = +1.85 for the use of video games on undergraduates' vocabulary acquisition. The study was based on only 40 undergraduates who were not randomly allocated to treatment conditions and there was no pre-test to confirm initial equivalence. In addition, the test instrument was developed

by the researcher who was not independent of the intervention. Another study in Chen et al.'s review (synthesised by Lewin et al.) reported that gaming in mobile learning enhanced children's learning of vocabulary (Sandberg, Maris & Hoogendoorn. 2014), with an effect size of d = 1.73. This study had attrition of over 30% and the treatment and comparison groups were in two schools. In other words, the number of cases allocated to treatments is only two. The effect sizes in Chen et al.'s review ranged from d = −0.024 to d = +2.301. By treating all the studies as of the same quality, the overall effect is likely to be skewed towards the weaker studies reporting huge effects. This is a widespread problem.

Reviews of the use of formative assessment (also known as assessment for learning or AfL) have suffered similar methodological problems. Formative assessment has been widely researched with studies reporting suspiciously large average effect sizes from d = +0.7 to +0.9 (Hattie 2018; Fuchs and Fuchs 1986). Evidence from previous meta-analyses is often based on pulling together large amounts of weak data. The most robust evidence in Hattie's review was by Kluger and DeNisi (1996) because it was considered to be 'the most systematic', it included studies that had at least a control group, measured performance, and had at least 10 participants. If this is the strongest evidence, it calls into question the reliability of all the other studies in the meta-analyses. Ten is a small number of cases. Kluger and DeNisi's (1996) meta-analysis was largely based on studies undertaken in controlled or laboratory conditions. Results may be different in real classroom conditions (See, Gorard, and Siddiqui 2016). Some studies were specifically for children with special educational needs, and children with behavioural, emotional and disruptive behaviour (Hattie 2009). For example, Fuchs & Fuchs' (1986) review was focused on 'handicapped' children (their term). And around one third of the studies reported negative findings anyway, suggesting that some feedback (which formative assessment is a part of) is ineffective under some circumstances (See 2020).

The seminal work by Black and Wiliam (2010), Inside the Black Box, was built on an earlier review by Fuchs and Fuchs (1986). All the studies in the review indicated substantial impact with average effect sizes between +0.4 and +0.7. Again, the average effects were based on studies for learners of all age groups (from age 5 to undergraduates) and across a range of subjects. Another review (Hopfenbeck and Stobart 2015) included 1,387 reports, most of which were small case studies (with perhaps one or two schools), while very few were large-scale or well-designed.

There is some evidence from a recent large-scale evaluation using a cluster randomised control design (Speckesser et al. 2018) that embedding formative assessment in classrooms has a small positive impact on pupil's overall Attainment 8 scores at GCSE (g = +0.1), but not for English and maths specifically. This is a large study involving 140 secondary schools in England and 25,877 pupils with low attrition. It is the strongest evidence we have so far of the small positive benefits of formative assessment for school aged children (in this case not involving EdTech). Kingston & Nash's (2011) review including only school-age children found only 13 studies with sufficient information for the calculation of effect size, and their findings suggest that formative assessment was more effective in English language arts than in maths or science (ES = +0.32; +0.17 and +0.09 respectively). Most of the studies were of poor quality in terms of the design. The average effect sizes may be slightly over exaggerated perhaps because the studies varied in

terms of quality. Overall though there is evidence that formative assessment (FA) using EdTech is promising.

All meta-analyses and hyper meta-analyses have reported wide variations in effect sizes depending on age group, the outcomes measured, instruments used, and the research design (e.g. whether it is randomised control, matched control or one-group pre-post design). They take no account of the bias introduced by attrition, treating a study with full response as equivalent to one with high dropout or missing data. See (2018) highlighted these issues where meta-analyses included both passive designs along with randomised control trials (RCTs), and no distinctions made between them. Jadad, Cook, and Browman (1997) raised the issue of discordant systematic reviews presenting conflicting results. Not all systematic reviews come to the same conclusions because of differences in research questions and selection criteria of included studies.

To overcome these issues, some researchers have simply excluded low quality studies (e.g. Cheung and Slavin 2012), such as studies with small samples, having no control group, or without established pre-intervention equivalence. Others (e.g. Sung, Chang, and Liu 2016) have excluded effect sizes that were unusually high, which seems too arbitrary. Pieper et al. (2012) identified a number of issues with such overviews of reviews. Key to these is the lack of critical appraisal of included studies. A number of reviews also rely on PRISMA/QUORUM (QUality Of Reporting Of Metaanalyses) an assessment tool, but these are guidelines for reporting systematic reviews that include items irrelevant to a review's quality. They do not assess the risk of bias in the individual studies being synthesised. Meta-analysing other reviews is therefore very problematic. Our review instead addresses these quality issues by looking at all studies, and passing each individual study through a quality check which takes into account research design, scale, attrition, measurement and other threats to validity. We then use the resultant weighting when judging the overall picture.

Formative assessment (FA)

Formative assessment, also known as Assessment for Learning (Black and Wiliam 2010), is feedback used as a classroom teaching pedagogy where teachers make regular, immediate and interactive assessment of students' learning. It usually begins with teachers finding out students' prior knowledge, and gaps in their knowledge. This can be in the form or oral questioning or quizzes to check for understanding and to monitor students' progress. Effective feedback is the heart of formative assessment (Clark 2003). The goal is to provide ongoing immediate feedback to help students understand what they already know, what they need to know and what they need to do to go further. However, not all feedback is formative. Feedback that simply says 'well done' or 'good work' is not formative as it does not provide information for the learner to use to make improvements. FA is, therefore, any feedback that provides the learner with information about their own learning and how to improve their performance (Gedye 2010).

Research has suggested that to be effective, feedback should be specific and related to the needs of the pupils (Hattie and Timperley 2007; Black et al. 2003; Crooks 1988). Previous studies also suggested that effective feedback has to be immediate (Anderson et al. 2001; Corbett and Anderson 1989; Kulik and Kulik 1988; Shute 2008) so that misconceptions can be corrected before they are assimilated and internalised. Providing

real-time feedback, that is adaptive to individual students' needs simultaneously to all students in the same classroom, is one of the biggest challenges in implementing effective feedback in real classroom conditions (See, Gorard, and Siddiqui 2016). It is not only time-consuming but often impractical (Wainer and Thissen 1993).

The use of digital technologies in facilitating formative assessment

Advances in technologies have now made it more for immediate and personalised feedback to be delivered to individual student adapted to their performance in real time feasible (Dzikovska et al. 2013, June; Liu et al. 2016). Increasingly, schools are using such technologies in the classroom to facilitate immediate feedback enmasse to students.

A number of digital based technologies have been developed in recent years that support children's learning. These commonly have built-in assessment features to provide immediate feedback to both teachers and students. These technologies are known by different names in the literature. The most commonly used ones are the student response system (Wang 2015), also known as audience response system (Pettit et al. 2015), classroom response system (Kortemeyer 2016), personal response system (Song, Oh, and Glazewski 2017) and learner response system (Wiggins, Sawtell, and Jerrim 2017). These are commonly used with clickers (Jones, Henderson, and Sealover 2009; Fuller and Dawson 2017).

Learner response system or clickers

Student response systems (SRS) or learner response systems are interactive assessment tools that collect student data – diagnostic, formative, or summative. SRS is based on the premise that timely and focused feedback facilitates learning. It provides real-time feedback to teachers about students' learning, and the data from the system allows teachers to provide immediate feedback to students. The data can be textual or numerical and collected via multiple choice or open-ended questions. The system maintains anonymity of student responders but also provides individualised feedback to teachers about students' learning. Some of these systems may also incorporate games to make learning more interesting (Jones et al. 2019; Md; Yunus 2019). The games themselves may not be formative, but studies have suggested that gamified SRS is more effective than non-gamified SRS in enhancing learning (Barrio, Muñoz-Organero, and Soriano 2016; Tan and Saucerman 2017) and better motivation and engagement (Wang, Zhu, and Sætre 2017).

SRS often involves the use of clickers, a little hand-held device with a number of buttons (e.g., labelled A-E) corresponding to answer choices to a question posed by the teacher (Bojinova and Oigara 2011). An example of a clicker is Kahoot. Kahoot is commonly used for quizzes and surveys. With Kahoot, the teacher projects questions on a big screen and the students answer these questions using their smartphones, tablets or computers.

Because of its simplicity and convenience, automated feedback based on multiple choice responses is widely adopted in formative assessment. There is growing research on the educational impacts of formative assessment using such automated

scoring technologies. But most current studies are conducted in higher education settings (e.g. Chen, Breslow, and DeBoer 2018; Gikandi, Morrow, and Davis 2011). For this review, our focus is on school-aged children from pre-school to post-secondary.

Because research suggests that the timing and the content of feedback are important factors in determining the effectiveness of FA, this review, therefore, will consider the efficacy of the content and timing of feedback delivered digitally either via desktops, ipads, tablets, mobile phones or other portable devices.

Computer-assisted formative assessment

There are also computer-assisted formative assessment systems where the feedback is built into the system, and learners respond to questions on their computer or digital devices and are instantly told if their answers are correct or not. There are several variations to this. Learners may simply be told if their answers are right or wrong, or they may receive suggestions as to how they might rectify that mistake. Some of these suggestions may be simply pointing out the mistakes; others are more elaborate with explanations (e.g. Maier, Wolf, and Randler 2016; Máñez, Vidal-Abarca, and Martínez 2019). Research has been conducted comparing the relative effectiveness of these variations in feedback types. In addition, teachers receive feedback on how each learner is progressing – for example, how long they take to answer each question, how many attempts they have made, and where they are going wrong. Some platforms have integrated diagnosis and formative assessment which are adaptive to the learners' progress (e.g. Wongwatkit et al. 2017).

As with the learner response system, some computer-assisted formative assessments platforms also used games to facilitate learning. Quizizz, Socrative and iSpring Learn are examples of gamified formative assessment system. Quizizz is a self-paced formative assessment, which evaluates students' understanding and provides a fun review. Socrative is an online assessment tool which allows teachers to provide in-class quizzes and also visualise and monitor student learning. Teachers can create their own multiple-choice questions or true/false questions and students respond in real time. The system provides feedback to students' responses in real-time. Like Socrative and Quizizz, iSpring Learn also allows teachers to create quizzes. Pictures, texts, and videos are used in the quiz and students compete to earn badges and points (Zainuddin et al. 2020). All three have a leaderboard showing children's progress.

Methods used in our review

The aim of this new review is to revisit and update the previous studies, and re-evaluate the evidence on the effectiveness of the use of EdTech in supporting formative assessment in the classroom. Unlike previous studies, we focus on single studies (not meta-analyses), and overcome the major methodological issues encountered in previous reviews by critically evaluating individual studies through the application of the Gorard 'sieve' (described below). The review, therefore, fills a gap in research in this area, and is so far the only review based on single studies, quality-weighted, and focused on school age children. Our main research question was:

Does the use of technologically assisted formative assessment improve students' academic and learning behavioural outcomes?

Because our initial literature review suggests that there are different types of technologically assisted formative assessment, the most common of which was the learner response system and its associated products like Kahoot, plickers, clickers and Socrative, we included these in our search terms. Questions for Learning is another popular product used in the UK and its software has been evaluated in a number of trials, which is why they are in the search syntax.

Search strategy

We began with a systematic search of sociological, educational and psychological electronic databases (Table 1) as well as Google Scholar, and Google (for unpublished and grey literature). We also followed up references in identified studies and existing reviews of literature, and used work that was known to us. The vast majority of the studies were found in the main electronic databases.

A list of keywords was developed relevant to the aims of the review:

'technology-enhanced formative assessment' OR 'assessment for learning' OR 'digital feedback' OR 'Assessing-to-Learn' OR 'Questions for Learning' OR 'hand-held learner response devices' OR 'hand-held devices' OR 'video-assisted instructional feedback' OR 'audio-assisted instructional feedback' OR 'computer-assisted instructional feedback' OR 'instantaneous feedback' OR Clickers OR Socrative OR Kahoot OR Plickers OR RecaP

AND evaluat OR interven* OR trial OR experiment OR review OR 'meta analys*' OR cause* OR effect* OR determinant OR 'regression discontinuity' OR 'instrumental variables' OR longitudinal OR 'randomi* control' OR 'controlled trial' OR 'cohort study' OR 'systematic review' OR impact*

Table 1. Database search outcomes.

Database/search engines	Number of studies picked up	Number exported to EndNote
Web of Science	1,177	42
Applied social sciences index and abstracts: ASSIA – ProQuest	448	None relevant
Scopus	214	18
ERIC – Ebscohost	1,123	26
Science Direct	19	19
Sage Journals	87	5
ProQuest Dissertations & Theses Global	146	3
PsycINFO – Ebscohost	717	12
British Education Index	174	3
JSTOR	2	2
Wiley Online Library	171	3
Handsearch (google and google scholar)	95,700	32
Total	99,978	165

AND child* OR school OR student OR teacher OR educat* OR K12 OR pupil

These were first applied to the electronic databases to test sensitivity in picking up relevant studies. The keywords included terms related to educational technology, formative assessment and young people's learning and wider outcomes, as well as terms relating to research designs that would be appropriate for testing a causal model, such as experiments, quasi-experiments, regression discontinuity and difference-in-difference. We did not set any date limits, to allow the search to be as broad as possible. The search was completed in July 2020. Slightly different search terms were used to adjust to the idiosyncracies of each of the search engines and databases.

Inputting the keywords above into the 11 databases/search engines resulted in close to 100,000 hits. An additional paper suggested by the reviewer of this paper was added. The records were first sorted by relevance. There is a filter function in electronic databases to rank records by relevance. These were then screened by eyeballing the title and abstracts until the next ten pages show no more relevant reports, when the screening stopped. As we screened down the page, we marked the relevant ones and put them in the folder (most databases have a folder to store marked records). These were then saved and exported to EndNote, a reference manager, for second round screening. In total, 165 were deemed relevant from reading the titles and/or abstracts.

Screening

As is normal in reviews involving multiple databases and keywords relating to technology, a large number of records were not relevant to the research questions but contained some of the keywords. To remove these, we eyeballed the entries looking at the title and abstracts and removed those that were clearly not relevant to the topic. We then screened for duplicates using the EndNote function. After excluding the duplicates, the full text of the studies retained were screened by applying the inclusion and exclusion criteria. This retained 61 studies (Figure 1). Five were subsequently excluded when it was clear that they did not present causal evidence. Fifty-five studies were eventually retained for in-depth data extraction. Of these two reported only attitudinal outcomes (Chan et al. 2019; Nation-Grainger 2017). These two are therefore not included in this paper, which focuses on attainment.

Inclusion criteria

Inclusion and exclusion criteria were determined prior to completing the searches and were applied after the initial screenings. Studies were included if they were:

(1) 1. Empirical research
2. About evaluation of education technology that supports formative assessment
3. Focused on school age children, from pre-school (age 5) to secondary (age 18) or K-12 in the US.
4. Conducted in mainstream, state-funded schools
5. About measurable academic and learning behavioural outcomes (e.g. attitude

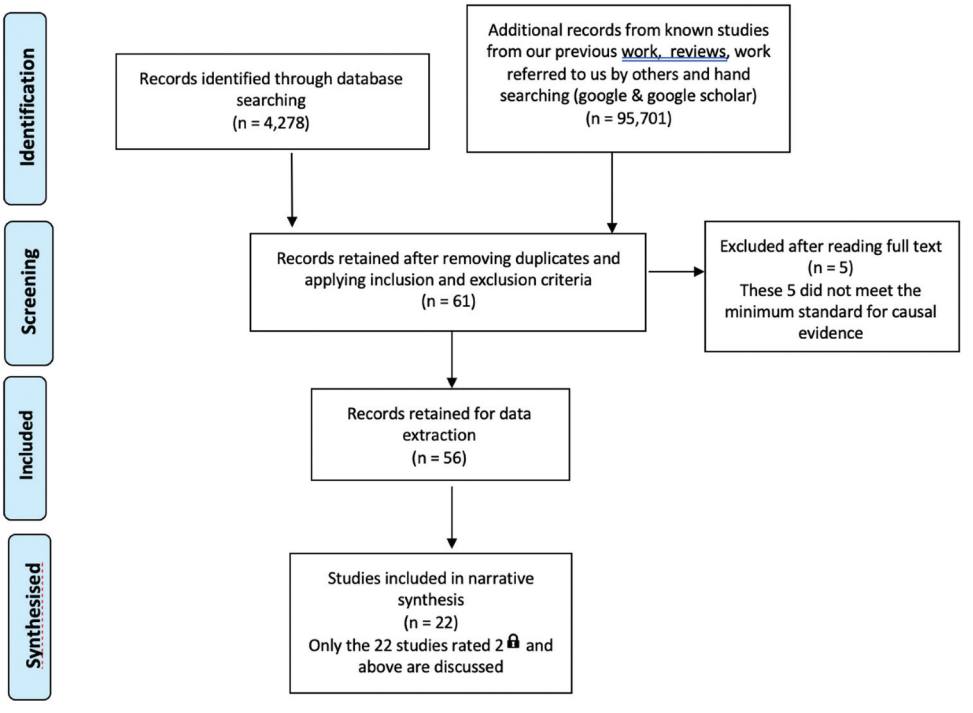

Figure 1. Flowchart of number of studies at each stage of the review.

and motivation towards learning)
6. Evaluation of impact of education technology

Exclusion criteria

Studies were excluded if they were:

(1) 1. Not relevant to the research questions
2. Not primary research
3. Not reported in English
4. Not a report of research
5. Descriptions of programmes or initiatives with no evaluation of the programme
6. Not about computer-assisted formative assessment
7. Studies that had no clear evaluation of outcomes
8. About non-academic outcomes not relevant to learning
9. Studies with non-tangible or measurable outcomes
10. Ethnographic studies and narrative case studies
11. Not about school-age children (i.e. adults or students in higher education)
12. Not relevant to the context of English speaking developed countries
13. Anecdotal accounts from schools about successful strategies
14. Promotional literature

Table 2. The quality appraisal 'sieve' (here for causal studies).

Design	Scale	Dropout	Outcomes	Other threats	Rating
Fair design for comparison (e.g. RCT)	Large number of cases per comparison group	Minimal attrition with no evidence that it affects the outcomes	Standardised pre-specified independent outcome	No evidence of diffusion or other threat	4🔒
Balanced comparison (e.g. Regression Discontinuity, Difference-in Difference)	Medium number of cases per comparison group	Some initial imbalance or attrition	Pre-specified outcome, not standardised or not independent	Indication of diffusion or other threat, unintended variation in delivery	3🔒
Matched comparison (e.g. propensity score matching)	Small number of cases per comparison group	Initial imbalance or moderate attrition	Not pre-specified, but valid outcome	Evidence of experimenter effect, diffusion or variation in delivery	2🔒
Comparison with poor or no equivalence (e.g. comparing volunteers with non-volunteers)	Very small number of cases pr comparison group	Substantial imbalance or high attrition	Outcomes with issues of validity and	appropriateness	
Strong indication of diffusion or poorly specified approach	1🔒				
No report of comparator	A trivial scale of study (or N unclear)	Attrition not reported or too high for comparison	Too many outcomes, weak measures or poor reliability	No consideration of threats to validity	0

15. Case study reports with no data evaluating outcomes
16. Data from developers' reports about the successful uptake of their programme

Quality assessment

Each included study was then assessed for the security of evidence using an appraisal tool known as the 'sieve' (Gorard 2018) based on five criteria (see Table 2):

(1) 1. Research design and fit to the study research question (e.g. for a causal question, whether it is an RCT with random assignment of cases, or matched comparison or longitudinal cohort study).
2. Scale of the study (smallest cell size in any substantive comparison)
3. Level of attrition/missing cases or data
4. Quality of outcome measurement (e.g. self-report or administrative data, standardised, independent or intervention-related assessment)
5. Other threats to validity (e.g. contamination, randomisation is subverted, conflict of interest)

This step is particularly useful in ensuring that the findings and conclusions made in the review are rooted in the quality of the available evidence. Pulling together poor data and biased evidence, however large the pool may be, is likely to mislead. Our review addresses the problems faced in most systematic reviews by weighting the evidence based on these

five criteria to ensure that the findings are weighted towards the results of the most robust studies.

Based on these criteria, each study is assigned a score using a padlock system between 0 (not of any value for our review), then 1🔒(the minimum standard to be given any weight, including some kind of comparison) and 4🔒(just about the best kind of causal evidence that can be expected from one real-life study). The latter are the most secure, meaning that the evidence is most appropriate for making causal claims. The 'sieve' reads from the top left corner starting with the design and moving right along the columns. For example, a large-scale randomised trial may start with a 4 and if there is noticeable attrition (perhaps resulting in observed and unobserved imbalance between the groups), then it will be rated 3 or even 2. It may then drop further if the test instruments are weak (i.e. they are designed by the programme developer, related to the intervention or based on participants' self-report). Ratings can only go down and not up. The ratings take no account of whether the intervention was deemed successful or not, or whether the report author claimed the intervention was effective. Where key information such as the amount of attrition is not reported, the piece is downgraded accordingly.

Of course, there are no objective criteria for deciding on any rating, or even on how many categories of ratings there should be. The table's cell descriptions do not include numeric thresholds. This is deliberate and leaves control in the hands of the research reviewer. Anyway things tend to settle down over the decisions in the four main columns. For example, the phrase a 'large number of cases' might be interpreted differently, depending upon the precise context, question or pay-off. There is also an interaction between the simple number of cases, their completeness, representativeness of a wider set of cases, their variance, and the integrity of the way they have been allocated to groups. 'A large number of cases' would certainly be in the hundreds, but there is no precise figure such as 400 that can be set, other than as a rough guide. An excellent study might have one case below whatever threshold is suggested (399), and a weaker one might have one more (401). Similarly, the importance of the amount of missing data depends on its balance between groups both numerically and in terms of the actual values, and the reasons for being missing. All decisions about research quality are judgements.

The Gorard 'sieve' is preferred specifically because it is easy to use, has shown high inter-rater reliability, and considers key factors that affect the validity of the findings. While there exists a few other quality assessment tools, e.g. the Maryland Scientific Scoring System (or SMS) and the PRISMA/QUORUM, they are not easy to use and omit key factors in their assessment – for example the scale of the study. SMS evaluates the robustness of research based on a 5-point scale ranging from 1, for evaluations based on simple cross-sectional correlations, to 5 for randomised control trials (RCTs). It considers the design of the study, but it does not take into account the scale of the study. In our review we found a large number of studies labelled as RCTs, but which randomised only two convenient classes or teachers, and involved very small number of cases. Such studies are usually conducted by researchers who are also the teacher using their own classes. According to SMS, these studies would still be rated 5* – which is absurd. QUORUM, on the other hand, is a set of guidelines for reporting systematic reviews that include items irrelevant to a review's quality. They do not assess the risk of bias in the individual studies being synthesised. Meta-analysing other reviews is therefore very problematic. The EPPI centre's review judged quality as whether the study answers

the review question, the relevance of the context of the study, the quality of execution in terms of accuracy, accessibility and clarity. Again, they are not specifically about the scientific rigour of the individual study.

To enourage inter-rater reliability and consistency, four members of the team reviewed and rated a sample of six papers, chosen because they were ambiguous in terms of relevance and their research design. Additionally, the team members were in consultation with each other throughout the data-extraction process.

Unlike previous reviews, we do not summarise the aggregated effect sizes as this may give a misleading impression about the efficacy of any programme. The key matter is whether the effect is positive or not, and how credible that effect is. The size of the apparent effect may assist in judging credibility, but can be misleading (as explained above). There are too few of each type of study to make averaging the effect sizes useful. For example, there are many variations in how these digitally assisted formative assessments programmes are used. Some compared different types of feedback (e.g. generic, contextualised and elaborated), some compared delayed and immediate feedback. While most were conducted in maths and reading classes, there were a small number that took place in social science lessons. The duration of the studies also varies widely. Therefore, we do not think it is desirable to average the overall effect size (Slavin 2020). However, we do report the effect size for individual studies where available, the direction of the effect (positive, negative or no change) and the strength of the evidence (i.e. how secure the finding is). Where the papers report the means and the standard deviation, the effect sizes are calculated by the reviewers using the difference in means between the comparison groups divided by the pooled standard deviation.

The findings

No studies were rated 4🔒, meaning that all had some identifiable limitations. There were a total of 56 studies relevant to the review, reporting 64 learning outcomes (relevant for the purposes of this paper). Of these, 22 were rated 2🔒or 3🔒in quality, and it is these that drive the review results.

As we reviewed the studies, we found that technologically assisted formative assessment software fell into two broad categories: those that we label as digital formative assessment, such as Questions for Learning and those that use a response system, such as clickers and Kahoot. For each type, we found that developers have also embedded games in them to make them more interesting and engaging. Therefore, we classified the programmes as with and without games – since the assumption by developers is that the use of games makes the programmes more effective. For this reason, we present the findings of this single review under four headings: digital formative assessment tool without the use of games; digital assessment tool with games; learner response system without games and gamified learner response system.

Digital formative assessment tool which provides feedback to teachers and pupils

Thirty-six studies with 43 distinct outcomes were concerned with the use of digital formative assessment tool that provides feedback to teachers and pupils. Seven of these outcomes were rated 3🔒, the highest rating in this review, and ten rated 2🔒(Table 3). One

study (2🔒) compared group and individual feedback and found the former to be more effective (Roschelle et al. 2010). The relatively large number of lower quality studies with apparently positive outcomes (22 here) is something we have observed before, and these studies in isolation can be largely ignored. What is less common is to find a disproportionate number of the stronger (3🔒) studies with positive results. This is promising. In this paper, we will discuss those rated 2🔒and 3🔒in more detail. Studies rated 1🔒and 0 will not be discussed in any detail here, as they do not add much to the overall evidence base, and do not change the substantive conclusions. A full list of these studies is available in Appendix A1.

Konstantopoulos, Miller, and Van Der Ploeg (2013) evaluated the impact of two assessment tools in Indiana for K-2 students. Positive effects were found for maths (ES = +0.19) and reading (ES = +0.12) based on the Indiana state's test and the Terra Nova test. The *mCLASS* provides formative assessments with diagnostic measures in literacy and numeracy. *CTB/McGraw-Hill's Acuity* is an online assessment tool in reading and mathematics for Grades 3 to 8, offering 30- and 35-item multiple-choice tests designed to be completed in a group setting. These assessments are both diagnostic and predictive. The study involved 59 schools that volunteered to apply *mCLASS* and Acuity (n = 2,000 grade 3 to 8 pupils), and were assigned to treatment (who have access to *mCLASS* and Acuity) or to business-as-usual control. Ten complete schools dropped out (attrition 17%), which is an important limitation – hence 3🔒.

An evaluation of an online assessment tool (ASSISTment) for maths homework (Murphy et al. 2020) reported positive effects for maths (ES = +0.22). There is an indication that the more frequently it is used the better the results. The intervention was found to be effective for lower performers. The ASSISTment platform allows teachers to assign homework, and students answer the questions on the platform. Students received automated feedback when they have completed and teachers receive reports on students' responses which they then use to adjust their teaching strategies. This was a waiting-list experimental study involving 46 schools, 87 teachers and 3,035 grade 7 students (age 12–13). Schools were randomised to ASSISTment or control – 3🔒.

Another digital formative assessment tool, Snappet, was found to have a positive impact on primary maths and children's attitude towards maths (Faber, Luyten, and Visscher 2017). Snappet allows students to complete the assignments using their own IT devices. Teachers and students receive immediate feedback from the system indicating whether their answers were correct. Snappet also offered adaptive assignments for each student, which were designed according to the previous performance level of each individual. Based on the feedback, teachers make decisions about the type of assignments for the individual pupil. The study, conducted in the Netherlands, included 97 primary schools randomly assigned to treatment conditions (40 treatment; 39 control;

Table 3. Summary of digital formative feedback and attainment outcomes (n = 43 outcomes).

	Positive impact (n = 35)	No impact/mixed (n = 5)	Negative impact (3)
3🔒	5	1	-
2🔒	8	2	1
1🔒	20	2	2
0🔒	2	-	-

n pupils = 1,808) over five months. The results indicated a positive effect on maths measured using a standardised test (ES = +0.39). The intervention appears to benefit boys and high performing pupils more than girls and lower performing pupils. Outcomes for maths performance was rated 3🔒.

In another randomised control evaluation of Snappet, but this time looking at the impact on spelling, Faber and Visscher (2018) included 69 primary schools (30 experimental and 39 control) with 1,605 pupils. Excluding the 10 schools that wanted only to do maths, attrition is 13%. Regression models revealed a small positive effect on spelling ($\beta = 0.05$) and spelling motivation ($\beta = 0.08$), controlling for pre-tests. The study was rated 2🔒because of the unclear attrition after randomisation, as there were an additional 10 schools that wanted to do only maths.

A randomised control trial of Questions for Learning (QfL) showed positive effects on children's grammar (ES = +0.16) but not writing (ES = 0) (Sheard and Chambers 2014; Sheard, Chambers, and Elliott 2012). QfL is a technology-enhanced formative assessment strategy which provides instantaneous feedback to teachers and pupils. However, because writing was not practiced as part of the intervention in the way that grammar was, the impact could be the result of practice. This was a 12-week trial involving 950 Year 5 pupils (age 9–10) from 42 primary schools in the north of England and Wales. Schools were matched by prior attainment and pupil characteristics before being assigned randomly to QfL or control – 2🔒.

In an earlier study, Sheard and Chambers (2011) conducted a 12- week randomised control trial in England involving seven schools in England (four experimental and three control) to test the effect of a Self-Paced Learning (SPL) strategy. Participants included 221 Year 5 pupils (109 experimental and 112 control). SPL provides pupils with one question after another on their screen, with an evolving graph on the teacher's screen for each child to show how they are answering and the duration for each answer. A positive effect was seen in maths (ES = +0.39). However, it has to be noted that outcomes were measured using pre- and post- tests developed by the evaluators based on the maths learning objectives in the National Framework for Year 5, but covering content addressed in the Learning Clip units used by the experimental group. There may, therefore, be practice effects. Since school-level randomisation effectively reduces the number of cases and as cases are the unit of randomisation, this and the fact that it used researcher-developed tests lower the evidence rating – hence 2🔒.

Siddiqui, Gorard, and See (2016) conducted an evaluation of Accelerated Reader (AR) in England involving 349 pupils in Year 7 (age 11–12) who had not achieved the expected Level 4 in their Key Stage 2 tests for English. Students were individually randomised to treatment conditions. Attrition was 2%. The intervention group of 166 pupils outperformed the 183 control pupils on the independent New Group Reading Test (ES = + 0.24). However, because AR is a multi-component intervention which includes explicit teaching, self-regulated reading and the use of technology for formative feedback it is difficult to say which of these components or combination of components drives the effect – 3🔒.

A slightly weaker evaluation of Accelerated Reader (AR)/Reading Renaissance (RR) conducted in the US (Ross, Nunnery, and Goldfeder 2004) reported mixed results on reading comprehension using the STAR reading test. This was an RCT involving 76 teachers and 1,665 pupils in 11 schools. Teachers were randomised to teach using AR/RR

or another commercially available reading programme. Attrition was 28%. The programme was found to be beneficial for children with learning disabilities (n = 978 in grade 3 to 6). Positive effects were found for all grades but bigger effects for lower grades (ES = +0.34 in third grade, +0.15 in fourth grade, +0.10 in fifth grade, and +0.07 in sixth grade). The study was graded 2🔒because of the high attrition and the fact that the STAR test was produced and marketed by Renaissance Learning as part of the AR programme.

Roschelle et al. (2010) compared the impact of Peer-Assisted Learning (TechPALS), a handheld technology that provides feedback in small groups with iSucceed Maths, a desktop product which provides feedback to individual students as they solve fractions problems individually. Two fourth grade classes (age 9–10) from three schools in the San Francisco Bay Area (n = 57 at School 1; n = 60 at School 2; n = 56 at School 3) were recruited. One class in each school was randomly assigned to TechPALS or control (iSucceed). The duration of the intervention was 12 days. In all three schools, students using TechPALS intervention (group work) made bigger gains than the control group (iSucceed) which used individual feedback (ES = +0.22) on a 29-item test of fractions. The study reported that low-scoring students at pre-test made bigger gains than students high-scoring students. As this could represent a regression to the mean, it is difficult to interpret. The results do not indicate if technology is beneficial, but rather that the use of technology is more effective in group feedback – 2🔒.

Fanusi (2016) examined the effect of ALEKS (Assessment and Learning in Knowledge Spaces), an online assessment tool on pupils' standardised maths test scores. ALEKS provides systematic instructions on maths concepts. The system applies artificial intelligence to decide the appropriate task for each individual student, and teachers can assign assessments on specific concepts. Participants were 294 grade 6, 7 and 8 (age 11–14) low-performing students from a rural middle school in Georgia. Students in the two support classes were not randomly allocated. Instead, pupils enrolled on ALEKS were matched by prior maths test scores, gender and socio-economic status with pupils enrolled in the traditional maths support class. Results showed that treatment pupils made less progress than control pupils (ES = −0.25). For those in the treatment group, there was a positive correlation between the number of ALEKS concepts completed and student's gain scores (r = 0.29). There is an issue with missing data as the analysis only included those who have test scores and only pupils who were enrolled in the maths support classes for the whole year were included. As the two classes of pupils were matched rather than randomly allocated, the two groups may be different in other unobserved characteristics (e.g. teacher effect, classroom climate or peer effect). It is therefore rated 2🔒(at best).

The finding is consistent with a meta-analysis of ALEKS conducted by Fang et al. (2019), which showed no benefit from ALEKS compared to traditional classroom teaching regardless of phases (i.e. secondary or postsecondary), whether the learning outcome was measured with standardised tests or instructor-designed tests or whether it was implemented as the principal or supportive instruction.

Maier, Wolf, and Randler (2016) explored the impact of computer-assisted formative assessment feedback on students' maths learning in northern Bavaria using software called Moodle to provide assessment and feedback in a 6-week cluster randomised trial. The computer-assisted assessment presents multiple-tier questions for each topic. The study was carried out in 10 secondary classrooms involving 261 grade 6 and grade 7 students (age 11–13). 14.2% of students either had no post-test data or retention test

results. Classrooms were randomly assigned to two treatment groups and a control group. The first received elaborated instruction-based feedback after formative assessments, the second received dichotomous verification feedback (simple information about the correctness of a response), and the control group only read appropriate texts instead of taking the formative tests and did not receive feedback. Students using verification codes achieved higher post-test scores in conceptual knowledge than those who used elaborated feedback. The author concluded that elaborated feedback is only helpful when students actually used it. No effect size was reported or can be computed. The test of conceptual knowledge were designed by the research team – 2🔒.

A study of the impact of a digital feedback tool in England found no impact on primary school children's maths (Sutherland et al. 2019). The Digital Feedback was delivered through a tablet application called Explain Everything. Teachers recorded videos giving their verbal feedback on pupils' work or took photographs of pupils' work both during lessons and outside of lessons. They then sent the videos or photographs to pupils through the application. The participants were 2,564 pupils in Year 4 and 5 (age 8 to 10) in 108 classes from across 34 schools in England. Fifty-six classes (1,103 students) were assigned to the treatment group and 52 (1,030 students) to the business-as-usual control. The maths outcome was measured using Essential Learning Metric (ELM) with KS1 test scores as the baseline assessment. The results showed that the control group had higher maths scores than the treatment group (ES = −0.04), and the treatment group displayed lower levels of engagement (ES = −0.09). The intervention also did not particularly benefit children eligible for free school meals in terms of maths performance and engagement (no average scores provided for FSM children). This study was rated 2🔒.

Shute, Hanson & Almond (2007) evaluated a computerised system of AfL involving 268 high school students randomised into 4 treatment conditions – simple feedback/adaptive sequencing, elaborated feedback/adaptive sequencing, elaborated feedback/linear sequencing and control (no assessment and no instruction). Results showed that elaborated feedback had positive effects on students with those using adaptive sequencing making more progress than those using linear sequencing. Both the simple adaptive and control group made negative progress. This suggests that elaborated feedback was the driver in enhancing learning. These were Algebra students and the test was geometry sequence. Control students received no instruction and tests. Since geometric sequences were not explicitly taught as part of the curriculum, comparison with control was not a fair test as they received no instruction in geometric sequences (the subject being assessed). The study was rated 2* because no mean scores and SD for pre-test were given, therefore not possible to calculate effect size. There was also no mention of number of students at post-test.

Zhu, Liu, and Lee (2020) evaluated an online formative feedback system integrated into a science curriculum module. The system provides immediate feedback in real time, responsive to students' progress. This dataset included 374 seventh to twelfth-grade students (age 12 – 18) from 22 classes taught by 8 teachers from 8 schools across the United States. Classes from each teacher were randomly assigned to either generic (10 classes and 145 pupils) or contextualised feedback (12 classes and 229 pupils). The generic feedback provided diagnostic information and improvement suggestions about students' explanations. Contextualised feedback, on the other hand, included details of

the context specific to each argumentation. Based on the feedback received, students made revisions to their answers. Among those who made revisions, the results showed that the generic feedback group performed better than the contextualised feedback group (ES −0.13), but more revisions were needed under the generic feedback condition to achieve similar score changes. The results indicated that the more revisions students made, the higher their final scores (ES = +0.73). This study does not evaluate the advantages of online formative feedback but does suggest that making revisions based on the feedback is an important element – 2🔒.

The remaining studies/outcomes are weaker in strength of evidence (rated 1🔒or below). Almost all reported positive effects for maths and reading. These studies were rated low on strength of evidence because of small samples (e.g. randomising two classes, one each to treatment and control conditions), very high attrition or missing data, no clear comparison, unclear or unreported samples and attrition, or lack of clear information about the data (e.g. missing pre-test data).

The stronger studies here (3🔒) suggest that the use of online or digital feedback has a beneficial effect for primary and secondary school children's maths and reading, but not writing. There is also some indication that the frequency of use is correlated with better performance, but the effect on lower performing pupils is unclear. There is no evidence of impact on other subjects (e.g. science, history and social studies) largely because of the relatively few and weak studies.

Digital formative assessment with game features

Multimedia tools such as video, animation and audio or games are sometimes incorporated in online formative assessment tools, to enhance motivation and improve interaction. We analysed such studies that embed these multimedia tools as a distinct group to see if the inclusion of this element contributes to learning and motivation (Table 4). Seven studies (8 outcomes) meeting our inclusion criteria evaluated such a programme, and all except one were rated 1🔒(see Appendix A2 for more details of the studies). The disproportionate number of positive lower quality studies is not unusual for any topic (Gorard, See, and Siddiqui 2017).

Song and Sparks (2019) compared the relative effectiveness of two types of feedback (answer-only versus explanatory feedback) using game-based formative assessment for the argumentation skills of 106 sixth and seventh graders. The programme includes some game features, such as interactivity, rules and constraints, challenges, goals, and immediate task-level feedback, which are displayed onscreen so that students may evaluate their ongoing performance and progress. Students using explanatory feedback made slightly bigger improvements in their argumentation skills than those who used answer-only

Table 4. Summary of studies of digital formative feedback with game features (n = 8 outcomes).

	Positive impact	No impact/mixed	Negative impact
3🔒	-	-	-
2🔒	-	1	-
1🔒	6	1	-
0🔒	-	-	-

feedback (ES = + 0.06). Students at the lowest two proficiency levels benefited from receiving the explanatory feedback (ES = +0.5), most students performed similarly across the feedback conditions, but highly proficient students performed worse on explanatory feedback compared to answer-only feedback. This was likely because one student scored particularly low, bringing the overall average down, demonstrating the volatility involved when using such a small sample. Although the study did not directly measure the impact on the state reading and writing tests, regression analysis showed that 37% of the variance in state test scores was explained by the post-test.

Overall, there is currently no good body of evidence that embedding gaming features in online formative assessment tools leads to any advantage (especially over the successful programmes without gaming as in Table 3). This contradicts the findings of the review, without quality control, by Wouters et al. (2013) who suggested that gaming features are particularly effective in language learning (ES = +0.66), less so for maths (ES = +0.17).

Clicker or learner response system

The little robust evidence on the Learner Response System (LRS) or Clicker suggests that it is ineffective or even harmful. The strongest study (with 2 outcomes), rated 3🔒(Table 5), suggests that there is no clear benefit on the academic outcomes of children using LRS (See Appendix A3 for the list of studies). **Wiggins, Sawtell & Jerrim's (2017)** evaluation of a Learner Response System (LRS) suggests that the intervention had no impact after one year and led to harm when used for two years. This is the strongest study on LRS, which used a cluster randomised control design. The LRS uses electronic handheld clicker devices that allow teachers and pupils to provide immediate feedback during lessons. The devices were to be used for 25 to 32 weeks a year. One cohort (Cohort A) used the device for one year and another for two years (Cohort B). The trial involved 6,572 pupils from 97 schools. Forty-nine primary schools were randomised to treatment and 48 to business-as-usual. School attrition was low at 2% and pupil attrition was 9% for maths and 12% for English. The results showed no effect on Cohort A's maths and reading (ES = 0) after controlling for pre-test (KS1 test scores). For Cohort B, there was a slight negative effect (ES = −0.08 for maths and −0.04 for reading).

The evidence from the 2🔒study with an attainment outcome is more positive. Zhu and Urhahne (2018) examined the impact on students' performance in maths, students' attitude to the technology and accuracy in teachers' judgement after five weeks. The sample included 459 sixth grade students from 20 classes across eight German middle schools. These classes were randomly assigned within the school to three treatment conditions: LRS with feedback, control class with regular maths instruction and no

Table 5. Summary of studies of the Learner Response System (n = 7 outcomes).

	Positive impact	No impact/mixed	Negative impact
3🔒	-	2	-
2🔒	1	-	-
1🔒	1	1	1
0🔒	1	-	-

Table 6. Summary of studies of the Learner Response System, with game features (n = 6 outcomes).

	Positive impact	No impact/mixed	Negative impact
3🔒	-	-	-
2🔒	-	-	-
1🔒	3	1	-
0🔒	1	1	-

feedback, and a diary group which received regular maths instruction, but the teachers were asked to reflect on their lessons. The results showed that the clicker group outperformed both diary (ES = + 0.44) and control (ES = +0.52) using the German 6th grade maths test. Since randomisation was at class level within each school, it means that around one class in each school was randomised to one of 3 conditions. This reduces the credibility of the findings as any differences between groups could be due to teacher quality or classroom dynamic.

Learner Response system/clicker with gaming

There is no good evidence that incorporating games with LRS enhances learning (Table 6). All studies had serious flaws in their design, such as having tiny samples (e.g. Lee et al. 2019), no comparison group, (e.g. Md. Yunus & Azmanuddin 2019), using intervention-related tests (e.g. Potter 2017; Yunus 2019; Tsihouridis, Vavougios, and Ioannidis 2017), or non-random allocation to treatment conditions (e.g. Sun & Hsieh). Most were small studies randomising two or three classes to each condition. Again, some of these weaker studies present results for non-attainment outcomes such as student engagement, motivation and anxiety. The full list of studies is available in Appendix A4.

Summary of review findings

The majority of studies that met our inclusion criteria involved primary and lower secondary pupils. Many reported positive results on learning outcomes. The stronger studies suggest that formative feedback delivered digitally can improve children's maths and reading, but not writing. The evidence relevant to science is not conclusive as there is only one medium rated study. Additionally, studies on science tend to focus on one topic in the curriculum.

Care has to be taken in interpreting the results because the implementation of these digital feedback tools varied considerably. Some involved generic feedback, some provided contexualised and elaborated feedback. Some feedback is delivered in real-time and some is delayed. How it is delivered may have implications for the results. Success may depend on the number of times students used an approach and the kinds of feedback involved. Fidelity is also a factor. Where students do not actually engage with elaborated feedback it is ineffective, or worse than marking and traditional instruction. Feedback may even be more effective in small groups (rather than individual), and when it is generic rather than precise.

Several of the studies suggested that digital feedback is more beneficial for low performing than for high performing students, but not always (Faber et al. 2018). There is no good evidence overall that adding gaming features or clickers, or both, is beneficial to attainment outcomes.

Conclusions

There is some promise that digitally delivered formative assessment can facilitate the learning of maths and reading for young school-age children. There is no evidence that it works for other school subjects or for older children. There is no evidence so far that learner response systems like Kahoot and Clickers work in enhancing children's academic attainment. There is also no robust evidence that embedding gaming features to these technologies make any difference to academic outcomes. This is largely because much of the research on the topic is so poor. Previous reviews may have suggested otherwise, but as discussed in our introduction, previous reviews rarely, if ever considered the trustworthiness of evidence based on research design and quality of data evaluated.

These new findings provide a caution against the current widespread investment in technology to improve school attainment. Schools and governments over the years have invested in such technologies and have, indeed, encouraged their use (and naturally have been encouraged to do so by the EdTech industry). However, while their use is widespread, the research on their use has not developed to the same extent. Schools wanting to use formative assessment, based on the evidence, can do so without the need for EdTech. EdTech may help or hinder, but is not in itself shown to be the solution to providing good formative assessment.

Before we promote the general use of programmes in schools it is important that their efficacy is rigorously tested in order to assess their effectiveness and scalability. Some of these technologies may even be harmful, leading to a delay in learning. Some may foster poor learning habits and there are also the opportunity costs. OECD (2015) highlighted the complexity of EdTech. Analysis of PISA data showed that students who used computers very frequently at school do worse in most learning outcomes than those who use them moderately, even after controlling for social background and student demographics. No obvious improvements in students' reading, mathematics or science were seen in countries that had invested heavily in information and communication technology (ICT) for education. Certainly it is not clear that teachers can be replaced by technology (Higgins 2015, September 15).

Replications of robust and large-scale studies are needed before we can ethically recommend the use of digital formative feedback in schools. This research is currently not happening enough. Research on the use of EdTech remain *ad hoc* and piecemeal, often with a partial motivation in terms of the funders or the researchers. The 2019 DfE plan for an EdTech testbed was shelved due to the coronavirus pandemic and funding was diverted to one-to-one online tutoring instead to support disadvantaged pupils who were the most likely to be adversely affected by missing school.

We should apply the same rigour and care in education research as we do in medicine. Large-scale stage 3 clinical tests on human subjects are necessary before medical products are administered to the population. Although children do not die from using poorly

conceived education technology, their one shot in education and life chances can be damaged if we do not use well-tested programmes or pedagogies which have shown to be effective. Such programmes do exist and are available, and only some involve the use of technology. We should refrain from being seduced by technology itself when we have so little evidence that it is effective.

Acknowledgments

This paper is based on research funded by Nesta and DfE.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Nesta and DfE.

Notes on contributors

Beng Huat See is Professor of Education Research at Durham University. She is a Fellow of the Academy of Social Sciences, Fellow of the Royal Society of Arts and the Wolfson Research Institute for Health and Wellbeing. Her research expertise is in the synthesis of research evidence and evaluations of education programmes and government policies.

Stephen Gorard is Professor of Education and Public Policy, and Director of the Durham University Evidence Centre for Education. He is a Fellow of the Academy of Social Sciences. His work concerns the robust evaluation of education as a lifelong process, focused on issues of equity and effectiveness.

Binwei Lu is a postdoctoral researcher in the School of Education, Durham University. Her research interest is in school effectiveness and equity, social justice and, social mobility and higher education access.

Lan Dong is a postdoctoral research associate in the School of Education, Durham University. Her recent work involves managing EdTech test-bed trials. Her interest is in culture-related phenomena.

Nadia Siddiqui is an Associate Professor in the School of Education, Durham University. She is a Fellow of the Wolfson Research Institute for Health and Wellbeing. Her work is concerned with understanding the patterns of poverty and inequalities and the role of schools in shaping a fairer society.

ORCID

Beng Huat See http://orcid.org/0000-0001-7500-379X
Stephen Gorard http://orcid.org/0000-0002-9381-5991
Binwei Lu http://orcid.org/0000-0002-3396-7635
Nadia Siddiqui http://orcid.org/0000-0003-4381-033X

References

Alcoholado, C., A. Diaz, A. Tagle, M. Nussbaum, and C. Infante. 2016. "Comparing the Use of the Interpersonal Computer, Personal Computer and Pen-and-paper When Solving Arithmetic Exercises." *British Journal of Educational Technology* 47 (1): 91–105. doi:10.1111/bjet.12216.

Anderson, T., R. Liam, D. R. Garrison, and W. Archer. 2001. "Assessing Teaching Presence in a Computer Conferencing Context." *JALN* 5 (2): 1–17.

Baker, E. L., National Center for Research on Evaluation, S., & Student, T. (2011). "Progress Report Year 4: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). The Development and Impact of POWERSOURCE[C]". CRESST Report 795. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED522826&site=ehost-live

Barrio, C. M., M. Muñoz-Organero, and J. S. Soriano. 2016. "Can Gamification Improve the Benefits of Student Response Systems in Learning? an Experimental Study." *IEEE Transactions on Emerging Topics in Computing* 4 (3): 429–438. doi:10.1109/TETC.2015.2497459.

Becta. 2009. *Home Access Programme: One Year On: Summary (December 2009)*. Coventry: Becta.

Bhagat, K. K., W.-K. Liou, J. Michael Spector, and C.-Y. Chang. 2019. "To Use Augmented Reality or Not in Formative Assessment: A Comparative Study." *Interactive Learning Environments* 27 (5–6): 830–840. doi:10.1080/10494820.2018.1489857.

Black, P., C. Harrison, C. Lee, Marshall, and Wiliam. 2003. *Assessment for Learning: Putting It into Practice*. New York: McGraw-Hill Education (UK).

Black, P., and D. Wiliam. 2010. "Inside the Black Box: Raising Standards through Classroom Assessment." *Phi Delta Kappan* 92 (1): 81–90. doi:10.1177/003172171009200119.

Bojinova, E. D., and J. N. Oigara. 2011. "Teaching and Learning with Clickers: Are Clickers Good for Students? Interdisciplinary." *Journal of E-learning and Learning Objects* 7 (1): 170–184.

Burns, M. K., D. A. Klingbeil, and J. Ysseldyke. 2010. "The Effects of Technology-enhanced Formative Evaluation on Student Performance on State Accountability Math Tests." *Psychology in the Schools* 47 (6): 582–591.

Byun, J., and E. Joung. 2018. "Digital Game-based Learning for K–12 Mathematics Education: A Meta-analysis." *School Science and Mathematics* 118 (3–4): 113–126.

Castillo-Manzano, J. I., M. Castro-Nuño, L. López-Valpuesta, M. T. Sanz-Díaz, and R. Yñiguez. 2016. "Measuring the effect of ARS on academic performance: A global meta-analysis." *Computers & Education* 96: 109–121.

Chan, C. H., A. S. Ha, J. Y. Ng, and D. R. Lubans. 2019. "The A+ FMS Cluster Randomized Controlled Trial: An Assessment-based Intervention on Fundamental Movement Skills and Psychosocial Outcomes in Primary Schoolchildren." *Journal of Science and Medicine in Sport* 22 (8): 935–940.

Chen, C. M., and M. C. Chen. 2009. "Mobile Formative Assessment Tool Based on Data Mining Techniques for Supporting Web-based Learning." *Computers & Education* 52 (1): 256–273.

Chen, M. H., W. T. Tseng, and T. Y. Hsiao. 2018. "The Effectiveness of Digital Game-based Vocabulary Learning: A Framework-based View of Meta-analysis." *British Journal of Educational Technology* 49 (1): 69–77.

Chen, X., L. Breslow, and J. DeBoer. 2018. "Analyzing Productive Learning Behaviors for Students Using Immediate Corrective Feedback in a Blended Learning Environment." *Computers & Education* 117: 59–74.

Cheung, A., and R. E. Slavin. 2012. "The Effectiveness of Educational Technology Applications for Enhancing Reading Achievement in K-12 Classrooms„ *A Meta-analysis*. Baltimore, MD: Center for Research and Reform in Education, Johns Hopkins University.

Cheung, A. C., and R. E. Slavin. 2013. "The Effectiveness of Educational Technology Applications for Enhancing Mathematics Achievement in K-12 Classrooms: A Meta-analysis." *Educational Research Review* 9: 88–113.

Chiu, Y. H., C. W. Kao, and B. L. Reynolds. 2012. "The Relative Effectiveness of Digital Game-based Learning Types in English as A Foreign Language Setting: A Meta-analysis." *British Journal of Educational Technology* 43 (4): E104–E107.

Chou, P. N., C. C. Chang, and C. H. Lin. 2017. "BYOD or Not: A Comparison of Two Assessment Strategies for Student Learning." *Computers in Human Behavior* 74: 63–71.
Chu, H. C. 2014. "Potential Negative Effects of Mobile Learning on Students' Learning Achievement and Cognitive load—A Format Assessment Perspective." *Journal of Educational Technology & Society* 17 (1): 332–344.
Chu, H. C., J. M. Chen, and C. L. Tsai. 2017. "Effects of an Online Formative Peer-tutoring Approach on Students' Learning Behaviors, Performance and Cognitive Load in Mathematics." *Interactive Learning Environments* 25 (2): 203–219.
Clarke, S. 2003. *Enriching Feedback in the Primary Classroom: Oral and Written Feedback from Teachers and Children*. London: Hodder and Stoughton.
Corbett, A. T., and J. R. Anderson (1989). "Feedback Timing and Student Control in the LISP Intelligent Tutoring System". In *Proceedings of the Fourth International Conference on AI and Education,* 64–72. Pittsburgh: Carnegie-Mellon University. http://act-r.psy.cmu.edu/wordpress/wp-content/uploads/2012/12/168corbett_and_anderson_feedback.pdf
Cosgrove, J., D. Butler, M. Leahy, G. Shiel, L. Kavanagh, and A. M. Creaven. 2014. *The 2013 ICT Census in Schools–main Report*. Dublin: Educational Research Centre.
Crooks, T. J. 1988. "The Impact of Classroom Evaluation Practices on Students." *Review of Educational Research* 58: 438–481.
DfE. 2019. *Realising the Potential of Technology in Education: A Strategy for Education Providers and the Technology Industry*. London: Department for Education.
DfES. 2006.Computers for Pupils. 2006-08. Guidance for LAs and Schools (Support Pack). London: DfES.
Dunham, V. K. (2011). "*The Impact of a Student Response System on Academic Performance* (Pp. 1-82)". Dissertation Thesis, South Carolina State University.
Dzikovska, M. O., R. D. Nielsen, C. Brew, C. Leacock, D. Giampiccolo, L. Bentivogli, and H. T. Dang 2013, June. "SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Embodiment Challenge." In *Second Joint Conference on Lexical and Computational Semantics (* SEM): Seventh International Workshop on Semantic Evaluation (Semeval 2013)*. Vol. 2. Atlanta, Georgia: Association for Computational Linguistics.
Faber, J. M., and A. J. Visscher. 2018. "The Effects of a Digital Formative Assessment Tool on Spelling Achievement: Results of a Randomized Experiment." *Computers & Education* 122: 1–8.
Faber, J. M., H. Luyten, and A. J. Visscher. 2017. "The Effects of a Digital Formative Assessment Tool on Mathematics Achievement and Student Motivation: Results of a Randomized Experiment." *Computers & Education* 106: 83–96.
Fang, Y., Z. Ren, X. Hu, and A. C. Graesser. 2019. "A Meta-analysis of the Effectiveness of ALEKS on Learning." *Educational Psychology* 39 (10): 1278–1292.
Fanusi, A. D. (2016). "*The Effect of ALEKS Math Support on Standardized Math Test Scores in Middle School*". Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2016-17334-232&site=ehost-live
Fuchs, L. S., and D. Fuchs. 1986. "Effects of Systematic Formative Evaluation: A Meta-Analysis." *Exceptional Children* 53 (3): 199–208.
Fuller, J. S., and K. M. Dawson. 2017. "Student Response Systems for Formative Assessment: Literature-based Strategies and Findings from a Middle School Implementation." *Contemporary Educational Technology* 8 (4): 370–389.
Gedye, S. 2010. "Formative Assessment and Feedback: A Review." *Planet* 23 (1): 40–45.
Gikandi, J. W., D. Morrow, and N. E. Davis. 2011. "Online Formative Assessment in Higher Education: A Review of the Literature." *Computers & Education* 57 (4): 2333–2351.
Gorard, S. 2018. *Education Policy: Evidence of Equity and Effectiveness*. Bristol: Policy Press.
Gorard, S., B. H. See, and N. Siddiqui. 2017. *The Trials of Evidence-based Education: The Promises, Opportunities of Trials in Education*. London: Routledge.
Gorard, S., B. H. See, and R. Morris (2016). "*The Most Effective Approaches to Teaching in Primary Schools: Rigorous Evidence on Effective Teaching*". Saarbrucken: Lambert Academic Publishing. Accessed online at: https://www.researchgate.net/publication/308395031_The_most_effective_approaches_teaching_in_primary_schools_rigorous_evidence_on_effective_teaching

Hattie, J. (2009). "Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement (Google eBook)".
Hattie, J. (2018). "*Hattie Ranking: 252 Influences and Effect Sizes Related to Student Achievement*". Retrieved from https://visible-learning.org/hattie-ranking-influences-effect-sizes-learning-achievement/
Hattie, J., and H. Timperley. 2007. "The Power of Feedback." *Review of Educational Research* 77 (1): 81–112.
Higgins, S. (2015, September 15). "Why Access to Computers Do Not Automatically Boost Students' Grades". *The Conversation*. Retrieved from https://theconversation.com/why-access-to-computers-wont-automatically-boost-childrens-grades-47521
Higgins, S., Z. M. Xiao, and M. Katsipataki. 2012. *The Impact of Digital Technology on Learning: A Summary for the Education Endowment Foundation*. Durham, UK: Education Endowment Foundation and Durham University.
Hopfenbeck, T., and G. Stobart. 2015. "Large Scale Implementation of Assessment for Learning." *Assessment in Education* 22 (1): 1–2.
Hwang, G. J., and H. F. Chang. 2011. "A Formative Assessment-based Mobile Learning Approach to Improving the Learning Attitudes and Achievements of Students." *Computers & Education* 56 (4): 1023–1031.
Instructure. 2017. *Driving Digital Strategy in Schools: The White Paper*. Salt Lake City: Canvas Instructure.
Jadad, A. R., D. J. Cook, and G. P. Browman. 1997. "A Guide to Interpreting Discordant Systematic Reviews." *Cmaj* 156 (10): 1411–1416.
Jones, S., D. Henderson, and P. Sealover. 2009. ""Clickers" in the Classroom." *Teaching and Learning in Nursing* 4 (1): 2–5.
Jones, S. M., P. Katyal, X. Xie, M. P. Nicolas, E. M. Leung, D. M. Noland, and J. K. Montclare. 2019. "A 'KAHOOT!' Approach: The Effectiveness of Game-Based Learning for an Advanced Placement Biology Class." *Simulation & Gaming* 50 (6): 832–847.
Keuning, H., J. Jeuring, and B. Heeren. 2018. "A Systematic Literature Review of Automated Feedback Generation for Programming Exercises." *ACM Transactions on Computing Education (TOCE)* 19 (1): 1–43.
Kingston, N., and B. Nash. 2011. „Formative assessment: A meta-analysis and a call for research.„ *Educational measurement: Issues and practice* 30 (4): 28–37.
Kluger, A. N., and A. DeNisi. 1996. "The Effects of Feedback Interventions on Performance: A Historical Review, A Meta-analysis, and A Preliminary Feedback Intervention Theory." *Psychological Bulletin* 119 (2): 254.
Koedinger, K. R., E. A. McLaughlin, and N. T. Heffernan. 2010. "A Quasi-experimental Evaluation of an On-line Formative Assessment and Tutoring System." *Journal of Educational Computing Research* 43 (4): 489–510.
Konstantopoulos, S., S. R. Miller, and A. Van Der Ploeg. 2013. "The Impact of Indiana's System of Interim Assessments on Mathematics and Reading Achievement." *Educational Evaluation and Policy Analysis* 35 (4): 481–499. doi:10.3102/0162373713498930.
Kortemeyer, G. 2016. "The Psychometric Properties of Classroom Response System Data: A Case Study." *Journal of Science Education and Technology* 25 (4): 561–574.
Kulik, J. A., and C. L. C. Kulik. 1988. "Timing of Feedback and Verbal Learning." *Review of Educational Research* 58 (1): 79–97.
Kulik, J. A., and J. D. Fletcher. 2016. "Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review." *Review of Educational Research* 86 (1): 42–78.
Lee, C. C., Y. Hao, K. S. Lee, S. C. Sim, and C. C. Huang. 2019. "Investigation of the Effects of an Online Instant Response System on Students in a Middle School of a Rural Area." *Computers in Human Behavior* 95: 217–223.
Lewin, C., A. Smith, S. Morris, and E. Craig. 2019. *Using Digital Technology to Improve Learning: Evidence Review*. London: Education Endowment Foundation.
Liu, O. L., J. A. Rios, M. Heilman, L. Gerard, and M. C. Linn. 2016. "Validation of Automated Scoring of Science Assessments." *Journal of Research in Science Teaching* 53 (2): 215–233.

Luckin, R., B. Bligh, A. Manches, S. Ainsworth, C. Crook, and R. Noss. 2012. *Decoding Learning: The Proof, Promise and Potential of Digital Education*. London: Nesta.

Lynch, S., G. Bielby, M. Judkins, P. Rudd, and T. Benton. 2010. *Evaluation of the Computers for Pupils Initiative: Final Report*. Coventry: Becta.

Maier, U., N. Wolf, and C. Randler. 2016. "Effects of a Computer-assisted Formative Assessment Intervention Based on Multiple-tier Diagnostic Items and Different Feedback Types." *Computers & Education* 95: 85–98.

Máñez, I., E. Vidal-Abarca, and T. Martínez. 2019. "Does Computer-based Elaborated Feedback Influence the Students' Question-answering Process?" *Electronic Journal of Research in Educational Psychology* 17 (47): 81–106.

Mertes, E. S. (2014). "*A Mathematics Education Comparative Analysis of ALEKS Technology and Direct Classroom Instruction*". Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2014-99210-460&site=ehost-live

Murphy, R., J. Roschelle, M. Feng, and C. A. Mason. 2020. "Investigating Efficacy, Moderators and Mediators for an Online Mathematics Homework Intervention." *Journal of Research on Educational Effectiveness* 13 (2): 235–270.

Nation-Grainger, S. 2017. "'It's Just PE' till 'It Felt like a Computer Game': Using Technology to Improve Motivation in Physical Education." *Research Papers in Education* 32 (4): 463–480.

Nikou, S. A., and A. A. Economides. 2016. "The Impact of Paper-based, Computer-based and Mobile-based Self-assessment on Students' Science Motivation and Achievement." *Computers in Human Behavior* 55: 1241–1248.

OECD (2015) "*New Approach Needed to Delivery on Technology's Potential in Schools*". (http://www.oecd.org/education/new-approach-needed-to-deliver-on-technologys-potential-in-schools.htm)

Ofsted (2011). "*ICT in Schools 2008-11. An Evaluation of Information and Communication Technology Education in Schools in England 2008-11*". London: Ofsted. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/181223/110134.pdf

Pettit, R. K., L. McCoy, M. Kinney, and F. N. Schwartz. 2015. "Student Perceptions of Gamified Audience Response System Interactions in Large Group Lectures and via Lecture Capture Technology." *BMC Medical Education* 15 (1): 92.

Pieper, D., R. Buechter, P. Jerinic, and M. Eikermann. 2012. "Overviews of Reviews Often Have Limited Rigor: A Systematic Review." *Journal of Clinical Epidemiology* 65 (12): 1267–1273.

Ponce, H. R., R. E. Mayer, V. A. Figueroa, and M. J. López. 2018. "Interactive Highlighting for Just-in-time Formative Assessment during Whole-class Instruction: Effects on Vocabulary Learning and Reading Comprehension." *Interactive Learning Environments* 26 (1): 42–60.

Potter, S. E. (2017). How integrating digital formative assessment impacts the learning of sixth-grade science students. Hamline University, School of Education Capstone doctoral thesis.

Reeves, J. L., G. A. Gunter, and C. Lacey. 2017. "Mobile Learning in Pre-kindergarten: Using Student Feedback to Inform Practice." *Journal of Educational Technology & Society* 20 (1): 37–44.

Rorabaugh, R. T. (2017). *Lowering the Cognitive Load through Differentiated Instruction Facilitated by Screen Casting in the Middle School Classroom* (Doctoral dissertation, Northcentral University).

Roschelle, J., K. Rafanan, R. Bhanot, G. Estrella, B. Penuel, M. Nussbaum, and S. Claro. 2010. "Scaffolding Group Explanation and Feedback with Handheld Technology: Impact on Students' Mathematics Learning." *Educational Technology Research and Development* 58 (4): 399–419.

Ross, S. M., J. Nunnery, and E. Goldfeder. 2004. *A Randomized Experiment on the Effects of Accelerated Reader/Reading Renaissance in an Urban School District: Final Evaluation Report*. Memphis, TN: Center for Research in Educational Policy, The University of Memphis.

Saffarian, R., and B. Gorjian. 2012. "Effect of Computer-based Video Games for Vocabulary Acquisition among Young Children: An Experimental Study." *Journal of Comparative Literature and Culture* 1 (3): 44–48.

Sandberg, J., M. Maris, and P. Hoogendoorn. 2014. "The Added Value of a Gaming Context and Intelligent Adaptation for a Mobile Learning Application for Vocabulary Learning." *Computers & Education* 76: 119–130.

Schleicher, A. (2015). "School Technology Struggles to Make an Impact". BBC News. https://www.bbc.co.uk/news/business-34174795

See, B. H. 2018. "Evaluating Evidence in Evidence-based Policy and Practice: Examples from Systematic Reviews of Literature." *Research in Education* 102 (1): 37–61.

See, B. H. 2020. "Why Is It Difficult to Get Evidence into Use? Chapter 7". In *Getting Evidence into Education: Evaluating the Routes into Policy and Practice*, edited by S. Gorard, 84-99, 2020. London: Routledge.

See, B. H., S. Gorard, and N. Siddiqui. 2016. "Teachers' Use of Research Evidence in Practice: A Pilot Study of Feedback to Enhance Learning." *Educational Research* 58 (1): 56–72.

Sheard, M., and B. Chambers. 2011. "Self-paced Learning: Effective Formative Assessment." In *Report on Achievement Findings*. York, UK: University of York.

Sheard, M. K., and B. Chambers. 2014. "A Case of Technology-enhanced Formative Assessment and Achievement in Primary Grammar: How Is Quality Assurance of Formative Assessment Assured?" *Studies in Educational Evaluation* 43: 14–23.

Sheard, M. K., B. Chambers, and L. Elliott. 2012. *Effects of Technology-enhanced Formative Assessment on Achievement in Primary Grammar*. York, UK: Institute for Effective Education.

Shute, V. J. 2008. "Focus on Formative Feedback." *Review of Educational Research* 78 (1): 153–189.

Shute, V. J., and S. Rahimi. 2017. "Review of Computer-based Assessment for Learning in Elementary and Secondary Education." *Journal of Computer Assisted Learning 33* 1: 1–19.

Siddiqui, N., S. Gorard, and B. H. See. 2016. "Accelerated Reader as a Literacy Catch-up Intervention during the Primary to Secondary School Transition Phase." *Educational Review* 68 (2): 139–154.

Slavin, R. (2020, April 16). "Cherry Picking or Making Better Trees?" https://robertslavinsblog.wordpress.com/2020/04/16/cherry-picking-or-making-better-trees/

Slavin, R., and D. Smith. 2009. "The Relationship between Sample Sizes and Effect Sizes in Systematic Reviews in Education." *Educational Evaluation and Policy Analysis* 31 (4): 500–506.

Slavin, R., and N. A. Madden. 2011. "Measures Inherent to Treatments in Program Effectiveness Reviews." *Journal of Research on Educational Effectiveness* 4 (4): 370–380.

Small, N. L. 2017. *The Effect of a Student Response System on Sixth-Grade Reading, English, and Language Arts Achievement*. Doctoral dissertation, Union University, ProQuest Dissertations Publishing. https://search.proquest.com/openview/33b71cf232826fe44d62045cedf4ebca/1?pq-origsite=gscholar&cbl=18750&diss=y

Song, D., E. Y. Oh, and K. Glazewski. 2017. "Student-generated Questioning Activity in Second Language Courses Using a Customized Personal Response System: A Case Study." *Educational Technology Research and Development*. doi:10.1007/s11423-017-9520-7.

Song, Y., and J. R. Sparks. 2019. "Building a Game-enhanced Formative Assessment to Gather Evidence about Middle School Students' Argumentation Skills." *Educational Technology Research and Development* 67 (5): 1175–1196.

Speckesser, S., J. Runge, F. Foliano, M. Bursnall, N. Hudson-Sharp, H. Rolfe, and J. Anders (2018). *Embedding Formative Assessment: evaluation report and executive summary*. London: Education Endowment Foundation.

Srisawasdi, N., and P. Panjaburee. 2015. "Exploring Effectiveness of Simulation-based Inquiry Learning in Science with Integration of Formative Assessment." *Journal of Computers in Education* 2 (3): 323–352.

Sun, J. C. Y., and P. H. Hsieh. 2018. "Application of a Gamified Interactive Response System to Enhance the Intrinsic and Extrinsic Motivation, Student Engagement, and Attention of English Learners." *Journal of Educational Technology & Society* 21 (3): 104–116.

Sung, Y. T., K. E. Chang, and T. C. Liu. 2016. "The Effects of Integrating Mobile Devices with Teaching and Learning on Students' Learning Performance: A Meta-Analysis and Research Synthesis'." *Computers & Education* 94: 252–275.

Sutherland, A., M. Broeks, M. Sim, E. Brown, E. Iakovidou, S. Ilie, H. Jarke, and J. Belanger. 2019. *Digital Feedback in Primary Maths: Evaluation Report*. London: Education Endowment Foundation.

Tan, P. M., and J. Saucerman. 2017. *Enhancing Learning and Engagement through Gamification of Student Response Systems*. Paper ID #18943. Columbus, Ohio, US: American Society for Engineering Education Annual Conference and Exposition.

The Scottish Government. 2015. *Literature Review on the Impact of Digital Technology on Learning and Teaching*. Edinburgh: Scottish Government.

Timmers, C. F., A. Walraven, and B. P. Veldkamp. 2015. "The Effect of Regulation Feedback in a Computer-based Formative Assessment on Information Problem Solving." *Computers & Education* 87: 1–9.

Tokac, U., E. Novak, and C. G. Thompson. 2019. "Effects of Game-based Learning on Students' Mathematics Achievement: A Meta-analysis." *Journal of Computer Assisted Learning* 35 (3): 407–420.

Topping, K. J., and A. M. Fisher. 2003. "Computerised Formative Assessment of Reading Comprehension: Field Trials in the UK." *Journal of Research in Reading* 26 (3): 267–279.

Tsai, F. H. 2013. "The Development and Evaluation of an Online Formative Assessment upon Single-Player Game in E-Learning Environment." *Journal of Curriculum and Teaching* 2 (2): 94–101.

Tsai, F. H., C. C. Tsai, and K. Y. Lin. 2015. "The Evaluation of Different Gaming Modes and Feedback Types on Game-based Formative Assessment in an Online Learning Environment." *Computers & Education* 81: 259–269.

Tsihouridis, C., D. Vavougios, and G. S. Ioannidis. 2017. "Assessing the Learning Process Playing with Kahoot–a Study with Upper Secondary School Pupils Learning Electrical Circuits." In *20thInternational Conference on Interactive Collaborative Learning*, Budapest, Hungary, 602–612. Cham: Springer.

Turan, Z., and E. Meral. 2018. "Game-Based versus to Non-Game-Based: The Impact of Student Response Systems on Students' Achievements, Engagements and Test Anxieties." *Informatics in Education* 17 (1): 105–116.

US Department of Education (2014) "*Learning Technology Effectiveness*". Office of Educational Technology. https://tech.ed.gov/wp-content/uploads/2014/11/Learning-Technology-Effectiveness-Brief.pdf

Vahdat, S., and A. R. Behbahani. 2013. "The Effect of Video Games on Iranian EFL Learners' Vocabulary Learning." *Reading* 13 (1): 61–71.

Van Der Kleij, F. M., R. C. Feskens, and T. J. Eggen. 2015. "Effects of Feedback in A Computer-based Learning Environment on Students' Learning Outcomes: A Meta-analysis." *Review of Educational Research* 85 (4): 475–511.

Vásquez, A., M. Nussbaum, E. Sciarresi, T. Martínez, C. Barahona, and K. Strasser. 2017. "The Impact of the Technology Used in Formative Assessment: The Case of Spelling." *Journal of Educational Computing Research* 54 (8): 1142–1167.

Wainer, H., and D. Thissen. 1993. "Combining Multiple-choice and Constructed-response Test Scores: Toward a Marxist Theory of Test Construction." *Applied Measurement in Education* 6 (2): 103–118.

Wang, A. I. 2015. "The Wear Out Effect of a Game-based Student Response System." *Computers & Education* 82: 217–227.

Wang, A. I., M. Zhu, and R. Sætre. 2017. *Does Gamification of a Student Response System Boost Student Engagement, Motivation and Learning?–An Evaluation of the Game-based Student Response System Kahoot*. Trondheim, Norway: NTNU Discovery.

Wang, K. H., T. H. Wang, W. L. Wang, and S. C. Huang. 2006. "Learning Styles and Formative Assessment Strategy: Enhancing Student Achievement in Web-based Learning." *Journal of Computer Assisted Learning* 22 (3): 207–217.

Wang, T. H. 2008. "Web-based Quiz-game-like Formative Assessment: Development and Evaluation." *Computers & Education* 51 (3): 1247–1263.

Wang, T. H. 2014. "Developing an Assessment-centered e-Learning System for Improving Student Learning Effectiveness." *Computers & Education* 73: 189–203.
Wiggins, M., M. Sawtell, and J. Jerrim (2017). *Learner Response System: Evaluation Report and Executive Summary*. London: *Education Endowment Foundation*.
Wongwatkit, C., N. Srisawasdi, G. J. Hwang, and P. Panjaburee. 2017. "Influence of an Integrated Learning Diagnosis and Formative Assessment-based Personalized Web Learning Approach on Students Learning Performances and Perceptions." *Interactive Learning Environments* 25 (7): 889–903.
Wouters, P., C. Van Nimwegen, H. Van Oostendorp, and E. D. Van Der Spek. 2013. "A Meta-analysis of the Cognitive and Motivational Effects of Serious Games." *Journal of Educational Psychology* 105 (2): 249.
Yang, K. H. 2017. "Learning Behavior and Achievement Analysis of a Digital Game-based Learning Approach Integrating Mastery Learning Theory and Different Feedback Models." *Interactive Learning Environments* 25 (2): 235–248.
Yunus, M., and M. Azmanuddin Bin Azman. 2019. „Memory Stay Or Stray?: Irregular Verbs Learning Using Kahoot.„ *Arab World English Journal* 5 (5): 206–219.
Zainuddin, Z., M. Shujahat, H. Haruna, and S. K. W. Chu. 2020. "The Role of Gamified E-quizzes on Student Learning and Engagement: An Interactive Gamification Solution for a Formative Assessment System." *Computers & Education* 145: 103729.
Zheng, L., X. Li, L. Tian, and P. Cui. 2018. "The Effectiveness of Integrating Mobile Devices with Inquiry-based Learning on Students' Learning Achievements: A Meta-analysis." *International Journal of Mobile Learning and Organisation* 12 (1): 77–95.
Zhu, C., and D. Urhahne. 2018. "The Use of Learner Response Systems in the Classroom Enhances Teachers' Judgment Accuracy." *Learning and Instruction* 58: 255–262.
Zhu, M., H. S. Lee, T. Wang, O. L. Liu, V. Belur, and A. Pallant. 2017. "Investigating the Impact of Automated Feedback on Students' Scientific Argumentation." *International Journal of Science Education* 39 (12): 1648–1668.
Zhu, M., O. L. Liu, and H. S. Lee. 2020. "The Effect of Automated Feedback on Revision Behavior and Learning Gains in Formative Assessment of Scientific Argument Writing." *Computers & Education* 143: 103668.

Appendix A1

Summary of studies on digital formative feedback (N = 36 studies)

	Reference	School phase	Types of outcomes	Direction of outcomes	Evidence rating
1	Konstantopoulos,Miller and van der Ploeg 2013	Primary and lower secondary (age 8 to 14)	Maths	Positive	3🔒
			Reading	Positive	3🔒
2	Murphy et al. 2020	Lower secondary (age 12–13)	Maths	Positive	3🔒
3	Sheard and Chambers 2014; Sheard, Chambers, and Elliott 2012	Primary (age 9–10)	Grammar	Positive	3🔒
			Writing	No effect	3🔒
4	Siddiqui, Gorard, and See 2016	Lower secondary (age 11–12)	Reading comprehension	Positive	3🔒
5	Faber, J. M., Luyten, H., & Visscher, A. J. 2017.	Primary (age 8–9)	Maths	Positive effect	3🔒

(*Continued*)

(Continued).

	Reference	School phase	Types of outcomes	Direction of outcomes	Evidence rating
6	Faber and Visscher 2018	Primary (age 8–9)	Spelling	Small effect	2🔒
7	Fanusi 2015	Middle school (age 11–14)	Maths	Negative	2🔒
8	Maier, Wolf, and Randler 2016	Secondary (age 11–13)	Maths	Positive effect in favour of simple (dichotomous feedback) vs elaborate feedback	2🔒
9	Roschelle et al. 2010	Primary (age 9–10)	Maths	Positive in favour of groupwork feedback vs individual feedback	2🔒
10	Ross, Nunnery, and Goldfeder 2004 (AR)	Primary to lower secondary (age 6 to 12)	Reading comprehension	Mixed results Positive effects for grades 1–3 but No effect on grades 4–6	2🔒
				Positive effects on at-risk pupils (pupils with learning difficulties)	2🔒
11	Sheard and Chambers 2011	Primary (age 9–10)	Maths	Positive	2🔒
12	Shute, Hansen & Almond 2007	Secondary	Maths (Geometry)	Positive effect of elaborated feedback	
13	Sutherland et al. 2019	Primary (age 8–10)	Maths	No effect (ES = −0.04) No effect on FSM	2🔒
14	Zhu, Liu, and Lee 2020	Secondary (age 12–18)	Science module on climate change	Positive effect of contextualised feedback and revisions	2🔒
15	Alcoholado et al. 2016	Primary (age 8–10)	Maths	Positive	1🔒
16	Baker 2011	Middle school (age 11–13)	Maths	Positive on one of 3 domains of maths	1🔒
17	Bhagat et al. 2019	Primary (age 9–10)	Science (environment)	Positive effect on knowledge of butterflies	1🔒
18	Burns, Klingbeil, and Ysseldyke 2010	Primary (age 3–11)	Maths	Positive Positive Also advantage of groups with >5 years compared groups with<5 years of use	1🔒
			Reading	Positive	1🔒
19	Chen and Chen 2009	Primary (age 9–11)	Maths	Positive, stronger effects for lower performing students	1🔒
20	Chou, Chang, and Lin 2017	Secondary (age 13–14)	English	Negative	1🔒
21	Chu, Chen & Tsai 2017	Primary (age 9–10	Maths	Positive effect of formative peer-tutoring	1🔒🔒
22	Chu 2014	Primary (age 10–11)	Social studies	Negative for test on social studies	1🔒
23	Hwang and Chang 2011	Primary (age 10–11)	Social studies	Positive on test of social studies	1🔒
24	Koedinger, McLaughlin, and Heffernan 2010	Secondary (age 12–13)	Maths	Positive More frequent use, bigger improvements	1🔒
25	Máñez, Vidal-Abarca, and Martínez 2019	Middle school (age 11–15)	Reading comprehension	Positive effect elaborated feedback partial η2 = 0.07	1🔒

(*Continued*)

(Continued).

	Reference	School phase	Types of outcomes	Direction of outcomes	Evidence rating
26	Mertes 2014	Middle school (age 11–14)	Maths District-developed test of concepts State comprehensive assessment	Mixed results Negative effect for grade 6 and 8 Negative for grade 7	1🔒
27	Nikou and Economides 2016	Secondary (age 16)	Science	Positive effect on physics	1🔒
28	Reeves, Gunther & Lacey 2017	Pre-K (age 4–5)	Print knowledge	Positive	1🔒
			Phonological awareness	Positive	1🔒
			Maths	Positive	1🔒
			Oral	Positive	1🔒
29	Rorabaugh 2017	Secondary (age 13–15)	Standardised writing assessment	Positive, but only temporary	1🔒
30	Srisawasdi and Panjaburee 2015	Secondary (age 14–15)	Science	Positive	1🔒
31	Wongwatkin, Srisawasdi, Hwang & Panjaburee 2017	Secondary (age 11–12)	Maths (measuring circle area)	Positive	1🔒
32	Zhu et al. 2017	Secondary (age 14–18)	Science module on climate change	Positive effect of revisions on argumentation	1🔒
33	Timmers, Walraven, and Veldkamp 2015	Secondary (age 13)	History – Regulation feedback on performance	Positive effect	0
34	Topping and Fisher 2003	Middle school (age 7–14)	Reading comprehension	Positive	0
35	Wang et al. 2006	Lower secondary (age 12–13)	Biology	No effect of FA (pre-test and learning styles stronger predictor of outcomes	1🔒
36	Wang 2014	Primary/Lower secondary (age 11–12)	Maths	Positive effect in favour of using gradual prompting	1🔒

Appendix A2

Summary of studies on digital formative feedback with game features (N = 7 studies)

	Reference	School phase	Types of outcomes	Direction of outcomes	Evidence rating
1	Song and Sparks 2019	Secondary (age 11–13)	Argumentation skills	Positive	1🔒
			Reading and writing	Mixed effect	2🔒
2	Tsai 2013	Secondary (age 14–15) (Tic Tac Toe quiz)	Science (knowledge on energy)	Positive effect (ES = +0.45) in favour of immediate elaborated feedback vs no immediate elaborated feedback	1🔒
3	Tsai, Tsai, and Lin 2015	Secondary (age 14–15	Science (knowledge on energy)	Positive effect of immediate elaborated feedback No difference between single and multiple player	1🔒
4	Vásquez et al. 2017	Primary (age 7–8)	Spelling	Positive effect in favour of group that used FA Tablet more beneficial than tablet IPC	1🔒
5	Wang 2008	Primary (age 10–11)	Biology	Positive effect in favour of game-based FA	1🔒
6	Yang 2017	Primary (age 10–11)	Social Studies (lesson on food cultures)	Positive effect in favour of corrective feedback vs regular feedback and pen and paper formative feedback	1🔒
7	Zainuddin et al. 2020	Secondary (age 15–16)	Geography (topics on landslides, volcanoes and flooding)	No difference between paper-based or gamified e-quizzes But self-paced FA (Quizizz) better than immediate feedback (Socrative) and no feedback (iSpring)	1🔒

Appendix A3

Summary of studies on Learner Response System/Clicker (N = 5 studies)

	Reference	School phase	Types of outcomes	Direction of outcomes	Evidence rating
1	Wiggins, Sawtell, and Jerrim (2017) Learner Response System	Primary	Maths	Mixed effects No effect after one year; negative after 2 years	3🔒
			Reading	Mixed effects No effect after one year; negative after 2 years	3🔒
2	Dunham 2011	Secondary	Maths	No effect	1🔒
3	Zhu & Urhahne 2019	Primary	Maths	Positive	2🔒
4	Small 2017	Middle school (age 10–13)	English & Language arts	Negative effect	1🔒
			Reading	Small effect	1🔒
5	Ponce et al. 2018	Sixth form	Reading		
	comprehension	Positive		0 Failed trial	

Appendix A4

Summary of studies on Learner Response System with games (e.g. Kahoot) (N = 6 studies)

	Reference	School phase	Types of outcomes	Direction of outcomes	Evidence rating
1	Lee et al. 2019	Secondary (age 12–13)	Earth Science	Positive effect in favour of Kahoot	1🔒
2	Sun and Hsieh 2018	Secondary (age 13)	English	No effect	1🔒
3	Tsihouridis, Vavougios, and Ioannidis 2017	Secondary (age 16–17)	Science (electrical circuits concepts)	Positive effect in favour of Kahoot	1🔒
4	Turan and Meral 2018	Secondary (12–13)	Social science Test of knowledge	Positive effect in favour of Kahoot vs Socrative	1🔒
5	Yunus 2019	Primary	Language – irregular verbs	Positive effect	0
6	Potter 2017	Middle school (age 11–12)	Science	No effect	0