STATE-OF-THE-ART REVIEW   **Review of Education** ▓BERA

Formative assessment and feedback for learning in higher education: A systematic review

Rebecca Morris[1] | **Thomas Perry[2]** | **Lindsey Wardle[3]**

[1]Department of Education Studies, University of Warwick, Coventry, UK

[2]School of Education, University of Birmingham, Birmingham, UK

[3]School of Education, Durham University, Durham, UK

Correspondence
Rebecca Morris, Department of Education Studies, University of Warwick, Coventry, UK.
Email: rebecca.e.morris@warwick.ac.uk

Abstract

Feedback is an integral part of education and there is a substantial body of trials exploring and confirming its effect on learning. This evidence base comes mostly from studies of compulsory school age children; there is very little evidence to support effective feedback practice at higher education, beyond the frameworks and strategies advocated by those claiming expertise in the area. This systematic review aims to address this gap. We review causal evidence from trials of feedback and formative assessment in higher education. Although the evidence base is currently limited, our results suggest that low stakes-quizzing is a particularly powerful approach and that there are benefits for forms of peer and tutor feedback, although these depend on implementation factors. There was mixed evidence for praise, grading and technology-based feedback. We organise our findings into several evidence-grounded categories and discuss the next steps for the field and evidence-informed feedback practice in universities.

KEYWORDS
evidence-informed practice, feedback, formative assessment, higher education, systematic review

Context and implications

Rationale for this study

To gain a better understanding of effective formative assessment and feedback approaches in higher education (HE). To promote a more evidence-informed approach to teaching and learning in universities.

Why the new findings matter

The findings highlight a small number of promising strategies for formative assessment and feedback in HE. They also draw attention to a lack of (quality) evidence in this area overall.

Implications for policy-makers and practitioners

Universities and their regulators/funders should be encouraging and supporting more, high-quality research in this important area. Researchers in the field also need to look to developing more ambitious, higher-quality studies which are likely to provide robust, causal conclusions about academic effectiveness (or other outcomes). Those involved in teaching and learning in university should use the findings to inform evidence-informed approaches to formative assessment and feedback and to challenge approaches which do not appear to have foundations in strong evidence. Students could be made more aware of teaching and learning approaches that are likely to support their academic progress.

INTRODUCTION

Formative assessment and feedback are fundamental aspects of learning. In higher education (HE), both topics have received considerable attention in recent years with proponents linking assessment and feedback—and strategies for these—to educational, social, psychological and employability benefits (Gaynor, 2020; Jonsson, 2013; van der Schaaf et al., 2013). On a practice and policy level there is widespread agreement that formative assessment and feedback should feature substantially within course design and delivery (Baughan, 2020; Carless & Winstone, 2019; OfS, 2019a). However, beyond this general expectation, it is less clear where the strength of evidence lies and what the most effective approaches and elements may be for HE students' learning (Boud & Molloy, 2013; Evans, 2013).

This systematic review examines the research evidence on the impact of formative assessment and feedback on university students' academic performance. It is the first international systematic review focusing on assessment and feedback in HE and presenting a comprehensive overview of causal evidence available in the field. Unlike other studies in this area, our review (a) employs a broad conceptualisation of formative assessment and feedback, including research across a range of different aspects of these pedagogical features, and (b) combines this with a rigorous quality appraisal process for identifying the most trustworthy, robust studies on which to base judgements about effective strategies.

There are currently over 200 million students enrolled in HE courses internationally, and this number is expected to continue to grow substantially in coming years (Calderon, 2018). Given this scale and the importance of feedback and formative education for learning, this systematic review has wide and significant implications for the field and for practice. We

indicate approaches and strategies where there appears to be some evidence for effectiveness while also highlighting the overall lack of high-quality, causal evidence available in this field. Implications of this for practitioners and policymakers seeking to work within an evidence-informed sector are also discussed.

This article proceeds as follows: in the two subsequent sections, we outline definitions of formative assessment and feedback and existing practices in HE relating to them. The methods section sets out our systematic review approach, including search terms, eligibility criteria, and details of the quality appraisal and analysis process. We then present summaries of studies presenting causal evidence, which we organise through categories grounded in the data and present through a narrative synthesis. Finally, we discuss the implications of our results for HE feedback and formative assessment research and practice, providing recommendations for the development of the field.

Definitions and types of formative assessment and feedback

There is no singular definition for either the terms 'formative assessment' or 'feedback'. Nevertheless, there is agreement that feedback is an integral element of a wider framework of formative assessment (Wiliam, 2018) and that both are concerned with the gathering and provision of information about a student's current performance or understanding to benefit students' learning. Black and Wiliam (1998), for example, describe formative assessment as including 'all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they [the students] are engaged' (Black & Wiliam, 1998, p. 8). As Sadler (1989) notes in earlier work, this transfer of information is not just between teachers and students. He argues that both peer and self-assessment can be important vehicles for providing feedback on students' existing performance and steps for moving forward.

It is this notion of addressing a 'gap' between students' current level of understanding and their desired level which typically forms a basis for definitions of feedback in education (Hattie & Timperley, 2007; Sadler, 1998). For some, using or storing this information to simply acknowledge a gap, however, is not enough; it must be utilised in a way to alter that gap, and ultimately have an impact on students' learning if it is to be called 'feedback' (Ramaprasad, 1983; Wiliam, 2011). For these intertwined processes of formative assessment and feedback to occur and work effectively, teachers are required to root them firmly within their pedagogical practices. Kluger and DeNisi (1996) in their seminal review, for example, stress that it is how students respond to or act on feedback that is more important than the type of feedback received. In order for this kind of response or action to happen, teachers therefore have to plan and embed opportunities for formative assessment and feedback activities into their curricula and teaching (Speckesser et al., 2018; Wiliam, 2018). Recent work by Carless and Winstone (2019) has indicated the importance of 'feedback culture' within HE. They describe the value of learning-focused models of feedback (as opposed to a one-way transmission model) whereby students are encouraged to actively involve themselves in engaging with and implementing the feedback they receive.

Hattie and Timperley (2007) identify four types of feedback, focusing on: (a) the task, (b) the process, (c) self-regulation, and (d) the individual. They argue that these have different purposes and variable impacts on students' learning. As a result of this they require different strategies for effective implementation. Most feedback is either verbal or written. Verbal feedback is frequently placed within the context of dialogue. From this perspective, feedback is seen as a 'move' within a dialogic teaching and learning approach (Hennessy et al., 2016; Perry et al., 2020). Feedback, for example, can range from a simple judgement of correctness, identification of a part of an answer that could be developed or improved, referring

back to prior contributions, and inviting opinions or ideas. Written feedback can take the form of corrections, marks, written comments, questions, targets and approaches designed to stimulate written dialogue. Written feedback is more typically focused on providing corrective and further information to develop student understanding rather than to inform teaching.

An increasingly important strand of educational thinking is emerging from cognitive science, which relates to understanding cognitive processes involved with memory and learning. Concepts such as working memory, long-term memory and cognitive load (Kirschner et al., 2006; Sweller et al., 2011) are influential in explaining how the human mind engages with, processes and retains information. Despite considerable interest in this work within the field of education, Wiliam (2018) points out that relatively few studies of feedback acknowledge these principles of cognitive science and instead tend to focus on shorter-term performance objectives linked to modes of feedback delivery rather than examining the deeper, longer-term processes of memory gain and learning (see Soderstrom & Bjork, 2015, for further discussion of the dissociation of learning and performance). Although much of the evidence base is currently derived from laboratory studies rather than 'real world' (i.e. ecologically valid) teaching and learning settings, cognitive science is providing a renewed emphasis on certain teaching strategies, including feedback strategies such as quizzing and frequent testing, which are rooted in evidence around recall and retrieval practice (Weinstein & Sumeracki, 2018). Cognitive science is likely to continue to offer theoretical bases and relevant evidence to develop understanding of feedback.

Evidence-informed formative assessment and feedback practice

Systematic reviews and meta-analyses, mostly conducted with compulsory school-age children, report relatively high average effect sizes ($d \approx 0.4–0.8$), albeit with large variation, ostensibly linked to a myriad of different forms of feedback, quality of implementation, and the teaching and learning context (EEF, 2018; Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Klute et al., 2017; Wisniewski et al., 2020). This evidence base tends to identify corrective feedback as more useful than praise, punishment or rewards for improving students' ability to learn new skills and complete tasks effectively (Hattie & Timperley, 2007; Kluger & DeNisi, 1996). Studies have highlighted that the more information included within feedback, the more beneficial it is and that the provision of comments is more helpful than simply sharing grades or marks (Hattie & Timperley, 2007; Wisniewski et al., 2020). Some reviews have examined the significance of the agents delivering the formative assessment and feedback: Klute et al. (2017) find that feedback directed by agents other than the student (i.e. a teacher or computer program) is more effective. Wisniewski et al. (2020) also tentatively highlight the effectiveness of peer feedback but note the small number ($n = 8$) of studies upon which they base this claim.

Some authors have suggested that written feedback may be more effective than oral feedback (Biber et al., 2011). However, the more recent meta-analysis by Wisniewski et al. (2020) found no evidence to support this claim. Unfortunately, research on the effects of written feedback is fairly limited and generally of low-quality: studies of written feedback at compulsory school level have concluded that although practitioners are frequently expected to spend extensive amounts of time providing detailed, written responses to their students' work, there is little evidence to suggest that it is effective in improving performance (Elliott et al., 2016).

As noted above, there are links between techniques derived from cognitive science and feedback. Research examining the impacts of quizzing and frequent testing is often rooted in the cognitive science literature, drawing upon theories of active recall and retrieval. The acts of recalling and retrieving information, often known as the 'testing effect' are believed

to support the long-term memorisation (and thus the learning) of that information (Dunlosky et al., 2013; Roediger & Karpicke, 2006). As a formative assessment tool, though, advocates of quizzes and testing point to benefits beyond remembering facts or key pieces of information. A 'feedback effect', they argue, can also support the development of conceptual understanding due to the opportunities that testing/quizzing provide to practise, develop and address errors or misconceptions when they occur (McDaniel et al., 2015; Vojdanoska et al., 2010). The extent to which this is possible depends upon the design and implementation of the quizzes/tests, and the contexts within which research is carried out. As with findings from cognitive science in general, much of the evidence on testing and quizzing is based upon trials conducted in laboratory settings. Although there have been some studies situated within 'real life' educational settings, there are few which are methodologically robust and even fewer involving post-compulsory educational institutions (Greving & Richter, 2018).

High-quality evidence, focusing on the impact of feedback on student academic performance in HE contexts is relatively thin compared to that found at school level. This raises questions about (a) what evidence at HE level reveals, and (b) the extent to which the evidence about compulsory school-age feedback applies to HE. A recent review of HE variables which influence student attainment, highlighted the potential value of different forms of formative assessment and feedback (Schneider & Preckel, 2017). Panadero and Alqassab's (2019) systematic review of anonymous peer feedback in HE included studies focusing on both school-age and higher-education level students, and tentatively suggests more positive impacts for those experiencing this approach at university.

Other reviews focusing more specifically on feedback in HE have tended to take a more conceptual and perspectives-based approach to understanding these issues. Evans (2013) set out to 'comprehensively explore the nature of assessment feedback within the specific and current contexts of HE' (Evans, 2013, p. 74), acknowledging also that the studies included within her review often draw causal conclusions where the research design or correlational findings do not warrant this. This study built upon earlier influential reviews such as that by Nicol and Macfarlane-Dick (2006), which sought to synthesise and reconceptualise the evidence in order to develop a more student-centred approach to feedback, moving away from it being viewed as merely an act of transmission for teacher to student (see also Carless & Winstone, 2019 for further discussion on this theoretical distinction). The authors present a model and seven principles of 'good feedback' for the development of student self-regulation of their performance. Although plausible and potentially useful, there is value in evaluating these broad principles, testing the impact of preferred and advocated strategies on student's *actual* progress and performance.

In summary, there is a considerable lack of research examining the impact of feedback and formative assessment on student learning in HE. To date, there has been no comprehensive study of this important area, presenting challenges for practitioners, institutions and policymakers who wish to adopt evidence-informed feedback and formative assessment practices. Our systematic review addresses this significant gap in the knowledge base and provides important recommendations for those working in HE settings and those researching in this field.

METHODS

The review addresses the following research questions:

1. What is the evidence of impact on student performance of formative assessment and feedback practices in HE?

2. What and how strong is the evidence of impact for different approaches to feedback?
3. What does the evidence suggest about principles for effective feedback and its implementation?

For the purposes of the systematic review, we considered educational performance to refer specifically to university students' attainment in assessments of academic performance. This may refer to their attainment in the subject that they were studying but could also include performance in other more generic academic skills, for example, essay writing where this has been assessed. We excluded other academic-related or wider outcomes such as attendance, progression, engagement with learning or enjoyment. Although these are important, and may well be linked to good assessment practice in HE, they were beyond our purview.

To identify all potentially relevant studies we searched the following electronic databases: Applied Social Sciences Index and Abstracts (ASSIA); British Educational Index, Educational Abstracts, ERIC (via Scopus); ProQuest dissertations and theses; ProQuest Central (Education, Psychology, Social Sciences, UK & Ireland); Social Sciences Abstracts; ACER; PsychInfo and PsychAbstracts; Ingenta Connect; and Web of Science. In addition, we carried out systematic searching using Google Scholar, retrieving the first 100 results following the searches with each of our criteria. Studies collated from additional hand searches, personal knowledge or that had been 'mined' from other reports, were also included at this early stage.

In line with our research questions, the search was for empirical studies that have examined the academic impact of formative assessment or feedback approaches in HE settings. Our key words cover the three relevant areas: first, the substantive topic—feedback or (formative) assessment; second, the setting/participants—HE and university-level students; and third, the causal nature of the research we were interested in—reflected via the design/methodological search terms. Different databases allow for and require different lengths, combinations and formats of search terms. Our general search terms, which we adapted to give the closest possible fit for every database were the following:

> *(Feedback OR assessment*) AND ("Higher education" OR "university student*" OR "college student*" OR "postgraduate" OR "undergraduate") AND (Trial OR experiment* OR "random*" OR RCT OR "regression discontinuity" OR "causal" OR quasi-experiment*)*

For each database, and where possible, we searched for these terms in titles, abstracts, and keywords. Searches were limited to publications in the English language and those published from the year 2000 onwards up until the search date of May 2019. After identification, all texts were downloaded into a reference manager. Following the removal of all duplicates, a total of 12,599 studies were included within this first stage. Screening of all titles was then completed to check for subject/topic relevance; following exclusion of irrelevant studies, we were left with 3290 records. The next stage of screening involved checking titles and abstracts and the application of our eligibility criteria to each piece (Table 1).

Following this process there were 188 studies which met the full eligibility criteria on inspection of full texts. Next, a process of information extraction for mapping was implemented to identify key details about each study such as geographical region, subject area, type and source of feedback/formative assessment and year. Alongside the overview data extraction process, we conducted a quality appraisal of each study, targeted at identifying causal evidence of impact. An evidence 'sieve' (Gorard et al., 2017) was used as a coding framework for this, requiring details on: study design, size, sample attrition, outcome quality, and threats to validity. Based upon these design and methodological elements, a 'quality' rating

TABLE 1 Eligibility criteria

Inclusion criteria	Exclusion criteria

• Focusing on feedback/formative assessment practices:
  ◦ Any medium—e.g. face-to-face/online
  ◦ Any format—e.g. marks/comments
  ◦ Any source—peer/self/tutor/ technology
  ◦ Any focus—feedback/'feedforward' / multi-direction
	• Not explicitly focused on feedback/formative assessment approach or practice (e.g. evaluation of an intervention where feedback/assessment is not a clear/distinct element).
	• Focused on summative feedback approaches with no formative element

• In Higher Education setting (level not name), including:
  ◦ HE course in Further Education college
  ◦ E.g. American 'college'
  ◦ Postgraduate and undergraduate
	• Study does not take place in higher education setting or with higher education-level students (e.g. school, sixth-form college)

• Study outcome produced via written output that is:
  ◦ testing a defined area of academic knowledge
  ◦ written e.g. English tests, written exams, dissertations, in-class quizzes
	• Outcome measure/output is not written (e.g. Performance in sport or music, speaking and listening skills in a foreign language, or an oral presentation)

• Includes a comparison group: ◦ at least two groups (i.e. intervention/control; pre/post intervention; within-subject design etc.)	• No comparison group (e.g. single group/cohort studies without a comparator)

of 1* (lowest quality) through to 4* (highest quality) was given to each study (see Gorard et al., 2017, for full details on the application of this tool). As a result, 27 studies were rated 3* and 1 was rated 4*; these 28 studies were retained for in-depth analysis in our narrative synthesis (see below). The remainder were mostly 2* (150) with a small number of 1* pieces (10). Our search terms had, by design, removed many studies that did not provide causal evidence, and would have otherwise been rated 1*. The full coding spreadsheet of included studies is available upon request from the authors.

In aiming to respond to our research question on the causal impact of formative assessment/feedback in higher education, we carried through only the 28 papers receiving a 3* or 4* quality rating for relevance and causal evidence for narrative synthesis.

Throughout each stage of the above process, checks were undertaken to ensure the quality, consistency and reliability in our judgements on the studies. During screening, each member of the research team took the same sample of titles/abstracts, comparing and discussing these with each other prior to continuation. For the quality appraisal stage, the authors checked inter-rater agreement by working with the same sample of studies to begin with. Borderline judgements were flagged for a second opinion, and these were discussed between the research team. Following this, the project leads also checked a random selection of studies and judgements prior to the final synthesis stages. Figure 1 provides a PRISMA diagram overview of the overall screening process.

RESULTS

An overview of the characteristics of the 188 eligible studies is provided in Table 2.

Table 3 also provides an overview of the quality ratings and the criteria used to determine these. Following this, we go on to present a narrative synthesis of the 28 highest-quality studies.

From these 28 studies, we identified the main topics and questions covered in each paper and then created five general thematic categories relating to the type, medium and delivery of feedback: (1) Content, detail and delivery, (2) Timing and spacing, (3) Quizzing and
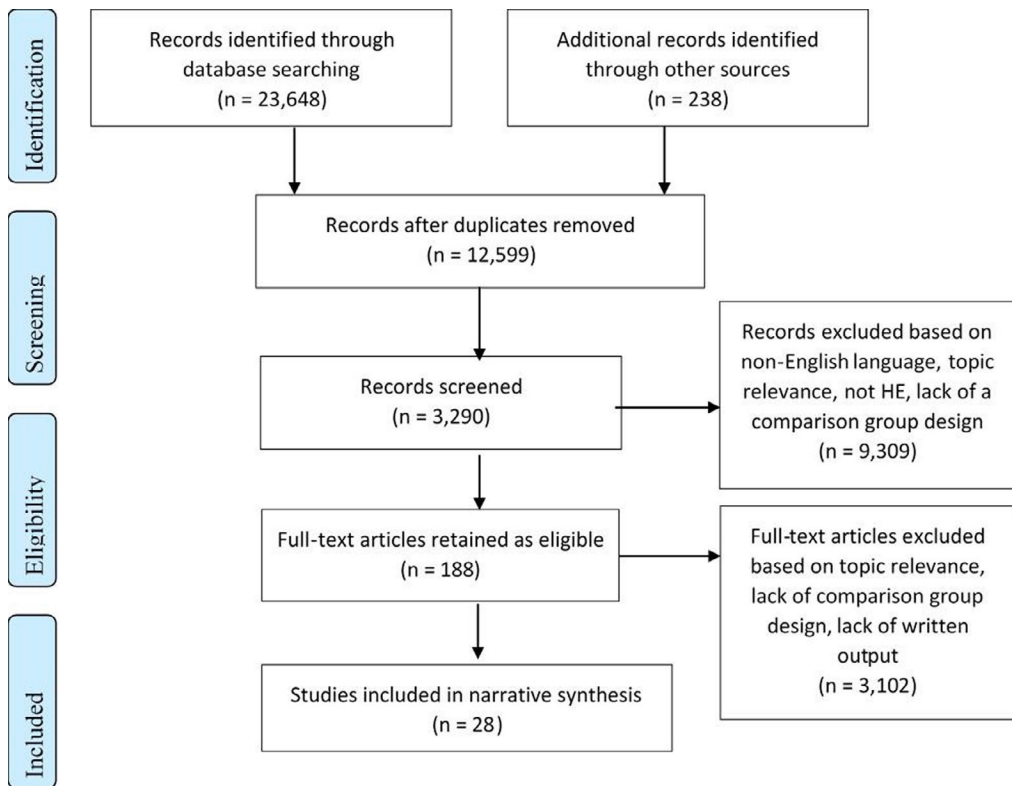
FIGURE 1 PRISMA flow diagram indicating number of studies included at each stage of the systematic review

Testing, (4) Peers, (5) Technology. See Appendix S1 for a table mapping all studies included in the detailed review against the general thematic areas. Below we provide a narrative synthesis of the studies in each thematic area, providing a description of each study and an overall summary of evidence within the theme. A small number of papers were identified as being relevant to more than one theme, but we report each within the theme to which it was most strongly aligned. Reporting each paper individually ensures that the full range of evidence from our relatively small number of remaining studies is presented openly and transparently for the reader. This approach also serves to highlight the breadth and diversity of studies here, and the challenges that this presents for developing a robust synthesis upon which to draw firm conclusions.

Content, detail and delivery

This section summarises high-quality studies focusing on the content and delivery of feedback and formative assessment. This includes research that examines a range of issues such as whether students receive feedback (or not), as well as the level of detail, amount and content of formative assessment tasks and feedback.

The strongest study in this section, and the only 4* rated piece within our review, is a natural experiment which examined the effect of providing feedback on past exam performance on future performance (Bandiera et al., 2015). The study used student data from Master's courses at a large UK university that were one year in length. Some departments provided students with feedback on their module exam performance (in the form of their

TABLE 2 Literature overview

Area	Category	Frequency	%
Feedback source	Mixture	26	13.8
	Peer	25	13.3
	Self	8	4.3
	Technology	53	28.2
	Tutor	76	40.4
Subject area	Arts and humanities	6	3.2
	(T)EFL	54	28.7
	Medical sciences	15	8.0
	Mixed or other	9	4.8
	Physical sciences, mathematics, engineering, technology	50	26.6
	Social sciences	54	28.7
Year	2001–2005	8	4.3
	2006–2010	37	19.7
	2011–2015	82	43.6
	2016–2019	61	32.5
Region	Africa	1	0.5
	Asia	24	12.8
	Central and South America	4	2.1
	Europe	46	24.5
	Middle East	27	14.4
	North America	80	42.6
	Oceania	6	3.2
Education Level	Postgraduate	10	5.3
	Undergraduate	175	93.1
	Across both or other	3	1.6

exam scores) immediately following these assessments across the year; a number of other departments did not do this, and only informed students of their exam performance at the end of their course (after all assessments had been completed). The researchers found that the provision of feedback had a positive effect on students' subsequent test scores with the mean impact corresponding to 13% of a standard deviation in test scores. The impact of the feedback was stronger for more able students and for students who had less information to start with about the academic environment, whereas no subset of individuals was found to be discouraged by feedback. This study indicates the importance and potential impact of providing timely information to students on their individual performance.

De Paola and Scoppa (2011) evaluated the impact of including an additional intermediate exam and providing students with information about their results prior to the final exam. Students in a control group took the final exam at the end of the module (without the additional mid-module exam). Participants were 344 students taking economics classes as part of a Business and Administration degree at a university in Italy. Half of the students were randomly allocated to the treatment group (mid-term exam) and half to the control group (final exam only). The results show that students undertaking the intermediate exam perform better both in terms of the probability of passing the exams and of grades obtained. High

TABLE 3 Quality rating of eligible studies (*n* = 188)

		Freq.	%			Freq.	%
Overall quality rating	1 (Lowest)	10	5.3	Positive findings	Unknown/unclear	6	3.2
	2	150	79.8		Negative	2	1.1
	3	27	14.4		Neutral	57	30.3
	4 (Highest)	1	0.5		Positive	123	65.4
Study size	Trivial/unclear	1	0.5	Attrition	Huge (50%+) or not reported	10	5.3
	Under 15 per comparison group (v. small)	11	5.9		High (30–50%)	9	4.8
	15–49 per CG (small)	114	60.6		Moderate (15–30%)	15	8.0
	50–99 per CG (medium)	38	20.2		Some (5–15%)	22	11.7
	100+ per CG (Large)	24	12.8		Minimal (0–5%)—no evidence of impact on findings	132	70.2
Design strength	Weak (flawed/biased comparison)	45	23.9	Outcome Quality	Uninformative	3	1.6
	Moderate (comparison made)	126	67.0		Poor—Issues of validity or appropriateness	6	3.2
	Strong (RCT/strong quasi-experiment)	17	9.0		Moderate—Tutor assessed, little QA or standardisation/implementation	135	71.8
					Strong—quality assurance, standardised, moderated	43	22.9
Threats	No evidence of limitations	47	25.0		V. strong—Standardised, pre-specified, independent	1	0.5
	Small limitations—will not affect results	82	43.6				
	Moderate limitation—might affect results	58	30.9				
	Serious limitations—probably substantial skew of results	1	0.5				

ability students appear to benefit more from the treatment. The design of the experiment also allowed the authors to understand whether this impact was due to 'workload division or commitment' effects or from 'feedback provision' effects. They found that the estimated treatment impact was due exclusively to the first effect, whereas the feedback provision had no positive effect on performance.

A number of studies within this section focus on the amount and/or type of feedback provided to students. This might include whether students receive feedback or not, the level of detail provided, or the use of written feedback and scores/grades. Lipnevich and Smith (2009), for example, examined the effects of providing no feedback versus detailed feedback to a large cohort of psychology students at a US university. Additionally, those provided with detailed feedback were either led to believe that it was provided by either the course instructor or computer generated. These conditions were also crossed with the receipt of a numerical grade (or not) and receiving a statement of praise (or not). All students were required to write a single-question essay at the beginning of their course. Detailed feedback on the essay, specific to individual's work, was found to be strongly related to student improvement in essay scores, with the influence of grades and praise providing more mixed results: receipt of a tentative grade depressed performance, although this effect was ameliorated if accompanied by a statement of praise. Overall, detailed, descriptive feedback was found to be most effective when given alone, unaccompanied by grades or praise. The perceived source of the feedback (the computer or the instructor) had little impact on the results.

Butler et al. (2008) examined the effect of immediate feedback compared with no feedback (until after completion of the post-test). Their experiment looked at feedback on regular online tests set as homework, rather than on a single formative task completed in class. Five sections of a mathematics course at a US university (total participants $n = 373$) were randomly allocated to either an immediate feedback or no feedback condition. Students in the immediate feedback group received information straight after completing each quiz. This meant that they could see their score and which items were answered incorrectly. Correct answers were not given to encourage the students to seek support with understanding their errors. The control group received no feedback (either scores or details of correct/incorrect responses) during the series of online quizzes; instead, they only found out this information after the end of the experiment. Results showed that students who received immediate feedback on quizzes had higher quiz and final test averages than those in the control group.

Heckler and Mikula (2016) investigated the levels of feedback complexity, studying the effects of 'knowledge of correct response' (KCR) feedback and 'elaborated feedback' (a general explanation) both separately and combined. Their study included 450 physics students learning about vector mathematics. Their findings indicated that elaborated feedback was most effective, especially for students with lower prior knowledge and lower course grades. In contrast, KCR feedback was less effective for these students. Combining both kinds of feedback also had no impact on students' performance compared to elaborated feedback alone. In a similar study, Petrović et al. (2017) also examined the impact of providing KCR or elaborated feedback, in comparison with a control group who received no formative assessments or feedback. Participants were three consecutive cohorts of students on a digital processing course at the University of Zagreb ($n = 70$—control group; $n = 34$—KCR group; $n = 35$—EF group). As the authors hypothesised, the results—based upon three summative assessments across the module—showed considerably higher performance for the two experimental feedback groups compared with the control group, who received no formative assessment. Further analysis also showed that those in the EF group performed better than those in the KCR feedback group in the summative assessments. Although there was no difference between the two experimental groups for the formative assessments, the authors suggest that the more detailed feedback is likely to have supported improved performance for the more complex tasks required as part of the summative assessments.

Two other 3* studies focused predominantly on the content of the feedback provided to chemistry students in a single US university. Scalise et al. (2018)'s experiment included two treatment groups: the first received additional conceptual questions in their online homework and the second received these questions plus differentiated answer feedback. Students receiving these interventions were compared with a business-as-usual group who received the usual online homework and feedback for the course. Both treatment groups showed increased gains in learning outcomes over the original comparison group. However, there were no differences between the two intervention groups, suggesting that the additional differentiated answer feedback may not have impacted performance any more than the use of conceptual questions on their own.

Like the above study which used additional conceptual questions to promote learning, Lee (2011) examined the use of learning strategy prompts and metacognitive feedback on students' outcomes. In this doctoral study, 261 undergraduate Education students were randomly allocated to three groups. One intervention group received learning strategy prompts, written statements which directed students to use different learning strategies when studying instructional material. A second intervention group received the learning prompts plus metacognitive feedback—information given to learners about their decisions regarding which cognitive strategies to use and how to use them. The third group acted as a comparison group. Two criterion tests measuring recall and comprehension served as post-tests. The study found that the participants who were given learning strategy prompts with metacognitive feedback scored significantly higher in the recall and comprehension tests after controlling for their prior domain knowledge. Those who only received the prompts (without the metacognitive feedback) scored no higher than the control group.

In a study with a different focus to those above, Mikheeva et al. (2019) investigated the role of politeness when giving instructions and feedback. In an online mathematics course at a German university, 277 students were randomly assigned to four groups: polite instructions and polite feedback (n = 64); direct instructions and polite feedback (n = 90); polite instructions and direct feedback (n = 57) and direct instructions and direct feedback (n = 66). Directness and politeness were characterised by factors such as numbers of words and vocabulary choices, and both instructions and feedback were provided online and in written form. Findings showed that politeness in instructions did not have an impact on outcomes, whereas receiving polite feedback did positively influence students' scores in the chapter tests and final post-tests.

As the above summaries highlight, there is considerable variation within this theme. The nature of these studies and their contexts are diverse; however, there are still some overarching conclusions that can be drawn. Perhaps unsurprisingly, we see evidence supporting the use of simple feedback (as opposed to no feedback) (Bandiera et al., 2015; Butler et al., 2008; Lipenvic and Smith, 2009; Petrović et al., 2017). In some settings, more detailed individual feedback is also shown to be effective, perhaps particularly for those with lower starting points in terms of attainment (Heckler & Mikula, 2016) and when completing more complex tasks (Petrovic et al., 2017). Evidence around the use of grades and praise is more mixed though (Lipnevic and Smith) and the study by De Paola and Scoppa (2011) indicates that including an additional assessment point may improve students' outcomes, but that this impact is not attributed to the feedback provided. There is little information provided about the influence of the source of feedback (i.e. via computer or instructor) although the findings from the studies here indicate that both can be effective. In terms of delivery though, Mikheeva et al.'s (2019) research suggests the importance of politeness in feedback provision. Work by Lee (2011) and Scalise et al. (2018) also points to potential promise for feedback activity which encourages students to spend time thinking more deeply about their work (e.g. via metacognitive strategies).

Timing and spacing

The timing of feedback provided following formative assessment activities emerged as one theme within the higher-quality studies. This tended to overlap either with issues raised in the section above (e.g. at what point feedback was provided) and with the studies focusing on quizzing/testing, where there was emphasis on frequent retrieval-based tasks to assess and feed back on learning. Here we discuss the two studies that foreground assessing the timing of feedback (immediate versus delayed) on student attainment.

Three studies consider the role of feedback timing during online formative assessment activities. These studies all examined the effect of giving feedback immediately (i.e. as students respond to each question item) or with a delay (i.e. following completion of the task). Van der Kleij et al. (2012) conducted a study with economics students at a university in the Netherlands. They randomly allocated students (n = 152) from nine classes to three different feedback condition groups. Following a formative assessment task involving an online, multiple choice question (MCQ) test, students either received immediate knowledge of correct response (KCR) and elaborated feedback; delayed KCR and elaborated feedback; or delayed knowledge of results (KR) but no additional feedback. An online summative assessment, used as a post-test, was administered immediately after the formative task. Findings indicate no significant difference between the feedback conditions and achievement on the post-test.

In a similar study, Gaona et al. (2018) also considered the impacts of immediate feedback provided on short-answer online quizzes. Their research, a quasi-experiment involving 5507 mathematics students across four university campuses in Chile, involved providing feedback on each question, including whether the response given was correct/incorrect, plus a step-by-step account of how to solve the question. One group of students received this feedback immediately after responding to each question (immediate) whereas the other group had to complete and submit the whole quiz before then receiving the feedback on each question (deferred). Findings from the study indicate that the Grade Point Average (GPA) was lower overall for students who received immediate feedback. However, the authors urge caution in interpreting this, pointing to the fact that students were allowed unlimited attempts at each quiz, and where they scored incorrectly on a question, they were likely to start the quiz again. Further analyses show that students spent longer on the immediate quiz feedback, took more attempts and achieved slightly higher maximum ratings. The authors suggest that these potentially positive outcomes need to be considered alongside the inefficiency and limited individual academic gain of this approach for students.

The findings from the studies above indicate a fairly unclear picture in relation to the value of immediate versus delayed feedback. This is echoed in a number of the 2* papers exploring issues of timing as well, signalling a need for further work in this area, and across different contexts and subject disciplines.

Quizzing and testing

Eight of the 28 higher-quality studies focused on quizzing or frequent testing, and its impact on student attainment. The majority of these include participants and content from science or maths-based subjects.

Four studies examine the impact of using quizzing/tests compared with either not using them or using alternative approaches. Peterson and Siadat (2009) evaluated the effect of frequent, cumulative, time-restricted multiple-choice quizzes with immediate constructive feedback on the achievement of mathematics students at a college in Chicago, America. Students were in groups which received either weekly or bi-weekly quizzes as formative

assessment, or in a control group which received no formative assessment. After four months, the results indicated that both quizzing groups performed better in their summative examinations than the control group. Doing the quizzes twice a week rather than just once appeared to have no additional benefit in terms of performance. In a similar study by Domenech Blazquez de la Poza and Munoz-Miquel (2015), students of microeconomics in a Spanish university participated in 10 short, handwritten, in-class tests across the course of one semester. These were cumulative and alternated between MCQ and problem-based, essay tests. To receive immediate feedback, suggested responses were immediately given to students following the test and marks were made available to students on the day of the test. When compared with groups not participating in the frequent testing approach, the findings indicate stronger performance on the final module exam for the testing group (an increase of 9.7 percentage points when control variables were included in the regression).

Pennebaker Gosling and Ferrell (2013) report the findings from a quasi-experiment examining the academic performance of students taking daily online, in-class quizzes which provided immediate and personalised feedback. Psychology students (n = 901) completed 26 short (10 min, eight MCQ items) tests during one semester; these contributed to 86% of the final grade for the module. Student performance was compared with the same data for classes previously taught by the same instructor (n = 935) but who had not used the frequent quizzing approach. Instead, this comparison group had completed four longer, written exams spread through the course of the term. Findings indicate a somewhat mixed picture. Students in the frequent testing group received lower grades overall than their predecessors in the control group. However, the authors posit that this is at least in part due to inflated (upward curving of) grades given to these earlier cohorts. Further analyses, including comparing results from the same questions used year-on-year, suggest that the experimental group's grades were higher by 0.59 of a letter grade. Using this as a constant, they go on to argue that when factored in, students in the intervention group performed better in their final assessment and in other classes too. However, the challenges with the outcome measures do mean that these results need to be interpreted cautiously.

A recent doctoral study by Sartain (2018) examined the effect of frequent testing on the exam scores of undergraduate nursing students at a US university. Four cohorts of students (n = 440) were allocated to either quizzing or non-quizzing groups with two cohorts per group. The non-quizzing group were required to undertake traditional unit exams and a comprehensive final assessment; the quizzing group were required to complete these as well but also had the addition of quizzes as part of their required coursework. One cohort within the quizzing group received instructions and information about the value of quizzing; the other did not. Analyses suggest that quizzing is linked to a positive impact on both unit and final exam scores, and that this was particularly the case for lower and middle achievers. There was no difference in attainment between the quizzing group who received the additional information on quizzing and the group that did not. The authors argue, therefore, that quizzing is an effective tool to help improve students' grades, regardless of whether students are made aware of its benefits or not.

Dobson et al. (2015) examined the extent to which testing—along with the reading of material—promoted greater recall and improved performance. Kinesiology students (n = 88) studied information relating to skeletal muscles, varying by three levels of familiarity (familiar, mixed information, unfamiliar). All students used both the repeated reading approach (R-R-R-R) and the read-test approach (R-T-R-T). The first studying strategy required students to read through a set of information on muscles four consecutive times. The second strategy asked students to first read through the information for 2 min and then spend 2 min testing themselves (through free recall) and repeat this process once. During the testing portions of the R-T-R-T strategy, students were unable to see the muscle information. Participants used the two strategies to study six sets of muscles in a sequential order and during just one

studying session. Learning was evaluated via free recall assessments administered immediately after studying and again after a one-week delay and a three-week delay. Across those three assessments, the read-only strategy resulted in mean scores of 29.3, 15.2 and 5.3 for the familiar, mixed and unfamiliar information, respectively, whereas the testing-based strategy produced scores of 34.6, 16.9 and 8.3, respectively. The results indicate that the testing-based strategy produced greater recall immediately and with a three-week delay, regardless of the participants' level of familiarity with the muscle information.

Through two experiments at a US university, McDaniel et al. (2015) also examined the effects of different sequences of testing and studying. For the first experiment participants (n = 85) read a research methods text. Two days later they were either assigned to: a first condition that involved repeatedly restudying the material three times (SSS); a second condition where they engaged in a test-restudy-test sequence (TST); or a third condition where they were tested on the studied material three times (TTT). All participants then received a final test five days later. Findings showed that both the TST and TTT produced better final performance than the SSS condition; however, TST was not better than TTT. In the second experiment (participants n = 124), the TST condition was altered so that after the first test, correct/incorrect feedback was provided and the test and feedback were available during the study phase. With this protocol, TST produced better learning and retention than did TTT or SSS. The authors highlight the correct/incorrect feedback given to participants after the first test as the 'critical modifier' here. This provided students with guidance of which areas of study that they needed to revisit before the second test to improve their performance.

A study by Rezaei (2015) examines the impact of frequent quizzing, both on an individual and collaborative basis. The study included 288 research methods students at a university in California, America. It compared groups of students taking part in the course between 2009 and 2014, all taught by the same instructor but using different assessment methods. The first group (control) followed the traditional approach of a mid-term test, final exam and research project. In the second group, the instructor also provided short (20 item), open-book online quizzes after each lecture. The third group completed all of the same elements except they were encouraged to take their quizzes in pairs. The findings indicate that the regular quizzing had a substantial positive impact on final grades, compared to the no quizzing condition. The authors note that there appeared to be a positive short-term effect (through improvements in the quizzes) and a longer-term effect too, as evidenced in the end-of-term exam. The group allowed to take their quizzes in pairs also went on to perform significantly higher than both the control group and the individual quizzing group, highlighting the potential promise for this kind of collaborative learning.

In two experiments on an educational psychology course, Vogler and Robinson (2016) also examine the effect of collaborative formative assessment. Their team-based testing (TBT) approach allowed students to work together to develop a consensus around test responses in three separate tests, answering until they were correct. As a comparison the students took another three tests individually with feedback. Students were then tested on this content two weeks later and again after two months. Results indicated that the TBT students scored higher when retested two months later than those who took the test individually.

The studies summarised above indicate considerable promise for quizzing and testing approaches. Evidence is presented for the benefits of using quizzing/testing within HE classrooms. Moreover, the process of including tests in pre- and post-study content, as well as asking students to complete them collaboratively, also appears to be a promising approach. Quizzing and testing is one of the more prevalent areas of assessment and feedback research that we found through our review. Although only a small number of studies were rated as 3* and summarised in this section, it is worth noting that positive findings were apparent from a number of 2* studies too. This evidence adds to the broader picture regarding this approach and its impact, and supports the suggestion that quizzing/testing is a 'good bet' for

supporting student learning and attainment in HE. What is less clear from the studies here, however, is the mechanisms that might support the effectiveness of quizzes/low-stakes testing. We do not know, for example whether it is the act of participating in these activities (i.e. the process of retrieval and recall that they require) or the feedback provided as a result of them that impacts students' improved learning and outcomes (see e.g., Halamish and Bjork (2011) who conduct a series of experiments examining the former). We return to the question of testing and the role of feedback within this as an operative mechanism in the final section of the article.

Peers

This section focuses on formative assessment activities or feedback which involves students working together to understand and develop their learning. Our review found five 3* studies relating to peer assessment or feedback. These examined the impacts of engaging with different kinds of peer review or feedback activities, including the use of ratings and qualitative feedback, providing peer review training, and the provision of anonymous or identifiable peer review. Overall, the studies in this area point to some potentially promising findings for strategies to support students' academic attainment. We discuss each one in more detail below.

Xiao and Lucking (2008) conducted a quasi-experiment, examining the impact of peer assessment on students' writing performance on a foundation teacher education course. A total of 232 online and campus students were divided in to two groups: one received ratings (in the form of numerical scores) on different aspects of their peers' writing; the other group received ratings and detailed qualitative feedback. Using the interactive software available on the Wiki online platform, four students were designated to assess each student's assignment. Following this first round of peer assessment, students were advised to rework their drafts and resubmit them. A further round of scoring then took place with multiple students assessing each piece of work. Final grades (and those used as the outcome measure of the trial) were awarded by instructors. Prior to the written tasks and assessment process, all students received a short briefing on peer assessment and the opportunity to practise scorings. Findings indicate that students in the scoring plus detailed written feedback group gained a small but significant improvement in their writing compared to the group who just received peer scores.

In their subsequent doctoral study, Xiao (2011) sought to examine the effects of peer-assessment skill training on students' writing performance. A quasi-experimental design was employed and included 473 foundation education students. Students from the first semester of the course (Group A—Fall semester 2007) formed the comparison group; they completed tasks as usual, using peer assessment but with no in-depth peer assessment skill training. A second group (Group B—Spring semester 2008) received principle-based peer assessment training, including two weeks of instructions on this approach. Principle-based peer assessment focused on the rationale for the approach, assessment criteria, ways to give effective feedback and judge peer performance. A third group (Group C—Fall semester 2008) received target criteria peer assessment training, including two weeks of instructions. This involve the same as the principle-based approach but was more closely integrated into the course content, more linked to the major assignment and required students to do peer assessment skill-focused exercises outside of the classroom. Using a similar Wiki article approach as above, students' pre- and post-scores in each group were compared. Findings show that students in both Groups B and C (who received in-depth peer assessment training) outperformed those in Group A. There were no differences, however, between the two

intervention groups, indicating that the more course-focused target-based approach was no more effective than the more generic principles-based approach.

Zhang's (2018) study also considers the impact of peer feedback on writing performance. This doctoral study included 198 English major students in a Chinese university. Eight intact classes were randomly assigned to either receive traditional, instructor-led feedback (control group) or peer feedback, including training on how to use and generate peer feedback (intervention group). The four classes in the intervention group did not receive instructor feedback for the 15 weeks of the study. Students completed initial assessments and a number of draft tasks with requirements to improve these following feedback from either the instructor (control group) or their peers (intervention group). For the intervention group students, peer feedback was delivered both orally and in writing. The author notes a difference in writing ability and English language proficiency between the two groups at the outset. However, even when taking these variables into account, they conclude a greater improvement for the treatment group. Analyses indicated that the quality of feedback that students received from each other was associated with their subsequent final grades. This was particularly the case when students had the opportunity to reflect upon the feedback that they had received from peers. Although potentially promising, caution is urged here due to the high effect size in relation to the academic performance, perhaps influenced by the differences between groups at the outset plus the fact that this was a relatively short, intensive intervention.

A study by Crowe et al. (2015) tested the effect of in-class student peer review in a quantitative research methods course. Based upon four sections of a course, 170 students completed two sections which incorporated in-class peer review and two sections which did not. For the two sections with peer review, content scheduled for the days during which peer review was used in class was delivered through an online course management system. Although the peer review activities took place in class, with the tutor present, the authors do not describe students being explicitly trained in peer review approaches. The findings show that in-class peer review did not improve final grades or final performance on learning outcomes for the module. Nor did it affect the difference in performance between drafts and final assignments that measured student learning objectives. Crucially, the authors also note the substantial amount of time that in-class peer review took and which meant that class delivery/teaching time was reduced, having a potentially negative impact on students' ability to access and engage with the full module content.

A final 3* study by Lu and Bol (2007) considered the effect of anonymous versus identifiable online peer review on writing performance. Participants were 92 undergraduate freshmen in four English composition classes enrolled in the Fall semesters of 2003 and 2004. The same instructor taught all four classes and in each semester one class was assigned to the anonymous e-peer review group and the other to the identifiable e-peer review group. All other elements—course content, assignments, demands, and classroom instruction—remained constant. Students completed eight e-peer reviewed written assignments through the term. Those in the anonymous group received feedback from two unidentifiable peers; those in the identifiable group worked in groups of three and all reviewed the work of each other. In both groups, reviewers provided suggested scores for the work, completed some editing and made suggestions for improvement. The results from both semesters showed that students participating in anonymous e-peer review performed better on the writing performance task. These students tended to provide more critical comments per draft and slightly lower scores than their colleagues in the identifiable group.

High-quality evidence on the impact of peer assessment and feedback is fairly limited. This section, however, does highlight some promising findings and should be read in conjunction with the subsection above, which indicates the potential benefits of collaborative quizzing and testing. Providing training for peer assessment appears to be useful in terms of promoting attainment and Lu and Bol's (2007) study also indicates the possibilities for

anonymised peer feedback. In addition to the research discussed here, there were a further eighteen 2* papers focusing on peer assessment and feedback. These were largely small-scale studies and mostly, like the studies above, focused on improving students' writing, often in English as a Foreign Language or social science settings. The majority of these report some positive findings (n = 12 studies) and also highlight some other benefits such as student engagement. Again, this suggests some degree of promise, albeit the need to consider the complexities of implementing peer feedback effectively and the potential cost of substituting instructional time for peer feedback (Crowe et al., 2015; Evans, 2013).

Technology

This section describes the higher-quality (3*) studies which have an emphasis on the use of technology in providing feedback to students. Studies included here focus on issues related to web-based feedback compared to paper or face-to-face feedback; the use of different web-based feedback systems for providing personalised performance information; and the use of technology such as video podcasting for providing feedback. We acknowledge that these are not the only technology-related studies included in the review. The use of online approaches for formative assessment and feedback can also be found in each of the other subsections; however, the ones described in this section are those that foreground the technology use and where it is the technology itself that is being explicitly assessed for impact.

Mitra and Barua (2015) examined the impact of online formative assessment and feedback versus a paper version combined with face-to-face feedback. The authors conducted a quasi-experimental trial with two groups of medical students in a single Malaysian university. The control group (n = 102) undertook a single paper-based formative MCQ test relating to the musculoskeletal module of their course and received whole-group face-to-face feedback on their performance. The experimental group (n = 65) instead received three web-based formative MCQ tests across the same five-week module and received automated online feedback. Despite students in the experimental group appearing to do better in the formative tests, in the final summative assessment (taken by all students in the study), there was no difference in overall performance.

In a further study, Richards-Babb et al. (2018) examined the use of an adaptive web-based feedback system for setting and responding to chemistry students' homework. This system involved providing a more personalised approach to completion of homework tasks and tests, giving students response-specific feedback on their work. This approach was compared with a traditional-responsive system (also online) where students were required to work through the same set of questions in the same order, regardless of their current level of mastery in the subject. Feedback for this approach also emphasised the need to correct mistakes. Using propensity score matching (n = 6114 pairs) to create comparable groups, the authors compared the outcomes of those students in the adaptive-responsive cohorts with those in the earlier traditional-responsive cohorts. The findings indicate that the adaptive system increased the likelihood of achieving a higher final grade, particularly for students who had average or below average prior attainment. Despite these potentially positive results, an accompanying attitudes survey showed that students reported less favourable attitudes towards the adaptive system compared to the traditional-responsive approach. This highlights the potential trade-off that HE lecturers sometimes face: a strategy that may support increased learning is not necessarily going to be received positively by students, particularly perhaps if it requires additional work or effort. Similarly, it is not necessarily the case that approaches which focus on providing engagement and enjoyment will also provide the best opportunities to maximise learning.

Chen (2011) examined the impact of an online personalised diagnosis and feedback tool which provides information to students on their learning paths. Computer programming students at a university in Taiwan (n = 145) were randomly allocated to either an experimental group (n = 72) who received the personalised online system following the completion of a formative test, or a control group (n = 73) who received just their test scores and no further feedback or engagement with the web platform. The personalised feedback system uses an algorithm (known as Pathfinder) to give students detailed information on the 'knowledge pathway' taken during the test and indicates misconceptions that occurred during the test. Comparisons of post-test scores show a mean of 58.9 (std. dev. = 15.5) for the control group and 68.2 (std. dev. = 14.75) for the experimental group. However, the authors urge caution that despite this potentially promising result, the experiment focused on only a single episode of using the online feedback tool.

A final study in this section reports two experiments involving video-based feedback (Leton et al., 2018). The first experiment tested the impact of providing knowledge of correct responses (KCR) (i.e. ticks/crosses) coupled with more detailed video podcast feedback compared with just KCR. The second experiment then compared the KCR + video podcast condition with KCR + written feedback/explanations (i.e. text-based explanations for the questions/responses). Participants in the first experiment were 44 engineering students taking a statistics course at a university in Madrid, Spain. After attending one theoretical and one practical lecture, students completed an online MCQ test using the Siette web platform, and either received KCR or KCR+video feedback. Results indicated that those in the experimental group achieved higher results in the post-test assessment. However, by this point numbers of participants were small, with just 16 remaining in the intervention group and 19 in the control group. The second experiment, undertaken in the following year, included more students (n = 112), allocated to either KCR+video feedback condition or KCR + equivalent illustrated feedback (text-based explanations). The results showed no difference in post-test performance between these two groups and also no difference in students' attitudes towards the different feedback methods.

The findings here indicate a rather mixed picture in relation to the use of online or video-based feedback. Where there are positive outcomes, these are often caveated with implementation or methodological issues. Moreover, there are challenges relating to the extent to which any impact (positive or negative) is associated with the use of technology as a mode for delivering feedback or as a strategy for generating and providing formative assessment and feedback (as seen for example with online quizzing). There were a further 42 studies with a 2* rating, which use technology in some way for the provision of feedback; as with the studies reported above, however, findings from these are very mixed. The studies indicate a real enthusiasm for employing learning technologies for feedback provision but little in the way of strong theoretical or empirical grounds on which to test effectiveness. These issues, plus the heterogeneous nature of the various technology-focused studies, makes it difficult to draw any firm conclusions about the benefits of using these kinds of approaches to deliver formative assessment and feedback in university.

DISCUSSION

This review has examined the impact of formative assessment and feedback in HE learning. The study set out to understand and summarise the evidence for these strategies and their impact on student performance. We identified 28 robust studies providing satisfactory causal evidence to test a form or quality of formative assessment or feedback. In this section we discuss the findings from these studies and present conclusions on the strength of

this evidence, and potential implications for policymakers, practitioners and researchers in the field.

In line with previous research, the evidence from our review provides support for the use of formative assessment and feedback for promoting attainment in HE. This will be reassuring for those HE lecturers who seek to base their practice on evidence-informed approaches to teaching and learning. Yet, despite this unsurprising high-level finding, we still know relatively little about the types, modes and features of these approaches that are likely to be most effective. The studies included in this review point towards some potentially promising strategies, including, for example, quizzing/testing and peer feedback. However, the limited and patchy nature of the research, plus the lack of methodological robustness in many of the studies means that it is difficult to offer firmer conclusions. Of the 188 records included in the final extraction and quality rating processes, 126 are based upon very small sample sizes, usually based in a single department or with a single lecturer in an institution. Often the studies appear opportunistic in nature, rather than being designed deliberately and with methodological rigour as a central consideration. This is perhaps partly related to the nature of HE teaching and research responsibilities for lecturers, and possibly linked to the challenges of gaining funding for more ambitious trials. Nevertheless, to provide a stronger evidence base, our review points to the need for a much more systematic and scaled approach to examining these vital areas of teaching and learning within HE. We discuss the possibilities for this further in the section below.

The evidence relating to quizzing and testing appears to suggest that embedding these approaches as a way of retrieving knowledge and identifying misconceptions or errors (for both student and teacher) can be beneficial. Our study finds that the majority of studies reporting the use of these approaches are based in science or mathematics-based subjects. This is perhaps because often such strategies focus on the recall of 'facts' or key pieces of knowledge, often associated with more technical learning. There were no studies in this review, across any quality rating, which tested the use of quizzing/testing within arts and humanities subjects. In addition, the quizzing/testing approaches are arguably more tightly focused and easily defined or operationalised than some other formative assessment and feedback approaches. This makes them more straightforward and attractive for the kinds of causal designs that we were looking for in this review. But while there may be more studies focusing on these strategies, most with positive findings, the investigations are still limited in probing whether it is the quiz (i.e. the process of retrieval) or the feedback received as part of it, which is likely to be the mechanism supporting improved attainment. Although some of the studies (in this section, and across the systematic review as a whole) have strong theoretical foundations, many do not. Similarly, a number of the studies with 2* and 3* ratings have low ecological validity (e.g. laboratory studies), again making it difficult for HE lecturers to find rich evidence that is relevant to their own setting. Although it is certainly a promising area of research to inform teaching, there is a need to continue with developing a more comprehensive evidence base from which to work.

Through our themes, we have begun to piece together a framework based on causal evidence. As we note above, this is necessarily limited and partial due to the lack of research in this area and further empirical studies are needed. When we compare the extent of the evidence on HE to that at school level, we find considerable disparity. Reviews and meta-analyses of research involving compulsory school-age pupils strongly suggests the importance of formative assessment feedback for supporting student progress and attainment. These findings and the extent of the evidence upon which they are based is not reflected in the HE literature. This seems curious given the size of the sector, the great pressure on universities to innovate, and the fact that there is often money available for teaching and research initiatives. Unfortunately, though, what appears to happen—based upon the published work that we have assessed through this review—is the development of myriad

new strategies, which are then rationalised and advocated rather than rigorously piloted, tested and (if successful) scaled. The approaches are frequently not rooted in strong existing evidence, and often focus on outcomes other than academic progress, such as student satisfaction, enjoyment or engagement. Small-scale evaluations of these approaches are sometimes carried out by those who develop them (and therefore are invested in highlighting positive findings) but these are rarely designed with a view to being able to make strong causal claims which add and build on the existing knowledge base. Additional methodological issues also arise when we consider the measures used to assess student attainment and progress. The majority of the studies included here used tutor-devised assessments, rather than more standardised approaches. This is not particularly surprising given that university assessment more broadly tends to be designed and implemented by tutors; externally assessed standardised tests or exams, as we see in the school sector, are much less common. This has potential implications for the reliability and validity of the results obtained through the studies included here, and also perhaps makes it more challenging to run multisite trials with multiple universities using the same standardised pre/post-tests.

There are a number of key issues here. First is the extent to which those teaching in HE are expected to undertake and publish research themselves, and the support available for doing so. Work from the school sector has highlighted the benefits and possibilities of engaging teacher practitioners within the development of a more evidence-informed system (Churches et al., 2020). In the context of universities, where teaching staff are often required to carry out research, the studies described in this review are likely to be useful and potentially informative. But a tension lies in the fact that this approach is not conducive to developing a broader, stronger evidence base that can be relied upon to inform teaching and learning policy on a larger scale. We are of the view that HE teaching is better advanced through the identification, testing and development of a set of key principles for effective HE teaching and learning, which lecturers can master and contextualise (in relation to subject and institutional context) than the desire to develop novel, 'innovative' approaches and conducting small-scale studies of their impact. Indeed, the higher-quality studies reported in this review have largely focused on the fundamentals of teaching and learning such as detail, timing, quality and delivery of feedback; these studies, and this review, are an important first step to building this HE-level evidence base. As noted above, the growing body of work emanating from pure and applied cognitive science holds promise for developing and explicating this evidence base (Agarwal et al., 2012; Churches et al., 2020). There also appears to be great value in development, side-by-side, with the evidence-base for compulsory school age pupils, for feedback and formative assessment, and teaching and learning more generally. While these are very different contexts, many of the fundamentals—including the value of high-quality communication, relationships and subject knowledge—are likely to remain important.

One of the key differences between the contexts of universities and schools is the aims and purpose of teaching and learning. Put simply, schools are usually expected to prioritise young people's academic progress. Along with other aims such as promoting children's safety, well-being and social outcomes, they are measured using performance outcomes (i.e. exam grades) and are held to account based upon these. In HE, this is less the case. Student attainment is not used as the main measure of university 'success' and a wider range of factors including student satisfaction and progression are collated via instruments such as the National Student Survey (in the UK) and are included as component measures in the Teaching Excellence Framework or university league tables. Within the current quasi-marketised system of HE, high value is placed upon students' perceptions around student experience, course satisfaction and value-for-money (Furedi, 2011) as this is what is measured and used for accountability purposes (OfS, 2019a, 2019b). There is arguably little incentive or opportunity for developing teaching and learning strategies focusing on

improving academic outcomes. In the UK, assessment and feedback as specific areas have typically received lower scores from students compared with other areas of university life (OfS, 2019a). This has led to universities being encouraged to enhance provision in this area (Nicol, 2010; OfS, 2019b). Although this focus on improvement is to be welcomed, the tension here remains: universities are encouraged to improve students' *feelings* of satisfaction with these areas, rather than embedding approaches that may also contribute to academic progress and performance.

Finally, we return to the issue of evidence-informed teaching and learning. If universities (and teachers in them) wish to provide the best opportunities for their students to achieve and reach their academic potential, then it is vital that policies and practices are focused on evidence-informed approaches. The government, regulators (such as the Office for Students in England) and other strategic organisations in the sector could also take a stronger role with supporting this stance and by investing resources. Tools and resources could be developed, similar to the school-based *Teaching and Learning Toolkit* (EEF, 2020) in England or the *What Works Clearinghouse* (WWC, 2020) in the USA to inform staff of useful strategies. Moreover, training of university teaching staff should model and foster the use of evidence to inform practice. That is not to say that there should be a 'one best way' approach to teaching in HE: practitioner autonomy and professional judgement is an important element of teaching in the university sector. However, we do think that there is an argument for widely sharing and promoting effective practices that could enhance students' opportunities to learn. Crucially, though, we would also suggest that there is the need for more research evidence upon which to draw. Without this, being 'evidence-informed' is much more challenging as we do not know what the 'best bets' are (Elliot Major & Higgins, 2019) and have a limited pool of information to base decisions upon. National bodies such as the OfS and Universities UK could play a vital and pioneering role in promoting, commissioning and funding larger-scale, methodologically rigorous and independent research studies in key areas of teaching and learning. Universities could also be encouraged and incentivised to participate in these to engage both practitioners and students in the pursuit of evidence-informed practice and genuinely impactful research.

LIMITATIONS OF THE REVIEW

Although this systematic review is robustly designed and reports findings fully and transparently, and effectively synthesises results and conclusions on a number of key areas of formative assessment and feedback, like all studies of this kind it has limitations. The most significant relates to the parameters of the review and the fact that our search terms and inclusion/exclusion criteria for date, language, and research design could have resulted in useful studies—which may have contributed to our knowledge and understanding—being excluded. We also acknowledge the potential publication bias that is revealed via our review (Torgerson, 2006). Nearly two thirds of our eligible studies (*n* = 123) reported positive results whereas only *n* = 2 (1.1%) published negative outcomes. Despite seeking to minimise possible publication bias by including unpublished and 'grey' material, it would still appear that positive findings relating to feedback in HE are more likely to be shared. We take this into account when discussing the studies and drawing overall conclusions, particularly regarding the need for more high-quality, larger-scale trials in this area.

CONCLUSION

Those teaching in HE care about learning and the achievements of their students. Although formative assessment and feedback appears to be a valuable approach to supporting

student performance, at present not enough is known about the specific and most effective strategies to be used. Our review contributes to a strong moral and academic case for an evidence-informed approach to teaching and learning in universities. For this to happen, the HE sector should learn lessons from recent movements towards evidence-use in the compulsory schooling sector. Ambition and commitment are needed but we are optimistic that this could lead to a stronger research base for practitioners to work with, and improved learning opportunities and outcomes for students.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICAL APPROVAL

As this research is based on a systematic review of published studies, this is not applicable to our research.

DATA AVAILABILITY STATEMENT

The database used for the collation and coding of studies included within this review is available upon request from the authors.

REFERENCES

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review*, *24*(3), 437–448.

Bandiera, O., Larcinese, V., & Rasul, I. (2015). Blissful ignorance? A natural experiment on the effect of feedback on students' performance. *Labour Economics*, *34*, 13–25. https://doi.org/10.1016/j.labeco.2015.02.002

Baughan, P. (2020). *On your marks: Learner-focused feedback practices and feedback literacy*. York: Advance HE. https://www.advance-he.ac.uk/knowledge-hub/your-marks-learner-focused-feedback-practices-and-feedback-literacy

Biber, D., Nekrasova, T., & Horn, B. (2011). The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis. *ETS Research Report Series*, *2011*(1), 1–110.

Black, P., & Wiliam, D. (1998). *Inside the Black Box: Raising standards through classroom assessment*. London, UK: King's College School of Education.

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education*, *38*(6), 698–712.

Butler, M., Pyzdrowski, L., Goodykoontz, A., & Walker, V. (2008). The effects of feedback on online quizzes. *International Journal for Technology in Mathematics Education*, *15*(4), 131–136.

Calderon, A. (2018). *Massification of higher education revisited*. http://cdn02.pucp.education/academico/2018/08/23165810/na_mass_revis_230818.pdf

Carless, D., & Winstone, N. (2019). *Designing effective feedback processes in higher education: A learning-focused approach*. London, UK: Routledge.

Chen, L.-H. (2011). Enhancement of student learning performance using personalized diagnosis and remedial learning system. *Computers & Education*, *56*(1), 289–299.

Churches, R., Dommett, E. J., Devonshire, I. M., Hall, R., Higgins, S., & Korin, A. (2020). Translating laboratory evidence into classroom practice with teacher-led randomized controlled trials—A perspective and meta-analysis. *Mind, Brain, and Education*, *14*(3), 292–302.

Crowe, J., Silva, T., & Ceresola, R. (2015). The effect of peer review on student learning outcomes in a research methods course. *Teaching Sociology*, *43*(3), 201–213.

De Paola, M., & Scoppa, V. (2011). Frequency of examinations and student achievement in a randomized experiment. *Economics of Education Review*, *30*(6), 1416–1429.

Dobson, J. L., Linderholm, T., & Yarbrough, M. B. (2015). Self-testing produces superior recall of both familiar and unfamiliar muscle information. *Advances in Physiology Education*, *39*(4), 309–314.

Domenech, J., Blazquez, D., de la Poza, E., & Munoz-Miquel, A. (2015). Exploring the impact of cumulative testing on academic performance of undergraduate students in Spain. *Educational Assessment, Evaluation and Accountability*, *27*(2), 153–169.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58.

EEF (2018). *Feedback review – Part of the teaching and learning toolkit*. https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/feedback/

EEF (2020). *Teaching and learning toolkit*. https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/. Accessed 18 September 2020.

Elliot Major, L., & Higgins, S. (2019). *What works? Research and evidence for successful teaching*. London, UK: Bloomsbury.

Elliott, V., Baird, J. A., Hopfenbeck, T., Ingram, J., Thompson, I., Usher, N., & Zantout, M. (2016). *A marked improvement? A review of the evidence on written marking*. London, UK: EEF.

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research*, *83*(1), 70–120.

Furedi, F. (2011). Introduction to the marketisation of higher education and the student as consumer. In *The marketisation of higher education and the student as consumer* (pp. 15–22). Abingdon, Oxon: Routledge.

Gaona, J., Reguant, M., Valdivia, I., Vasquez, M., & Sancho-Vinuesa, T. (2018). Feedback by automatic assessment systems used in mathematics homework in the engineering field. *Computer Applications in Engineering Education*, *26*(4), 994–1007.

Gaynor, J. W. (2020). Peer review in the classroom: Student perceptions, peer feedback quality and the role of assessment. *Assessment & Evaluation in Higher Education*, *45*(5), 758–775.

Gorard, S., See, B. H., & Siddiqui, N. (2017). *The trials of evidence-based education: The promises, opportunities and problems of trials in education*. Milton Park, UK: Taylor & Francis.

Greving, S., & Richter, T. (2018). Examining the testing effect in university teaching: Retrievability and question format matter. *Frontiers in Psychology*, *9*, 2412.

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 801.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Heckler, A. F., & Mikula, B. D. (2016). Factors affecting learning of vector math from computer-based practice: Feedback complexity and prior knowledge. *Physical Review Physics Education Research*, *12*(1), 010134.

Hennessy, S., Rojas-Drummond, S., Higham, R., Márquez, A. M., Maine, F., Ríos, R. M., García-Carrión, R., Torreblanca, O., & Barrera, M. J. (2016). Developing a coding scheme for analysing classroom dialogue across educational contexts. *Learning, Culture and Social Interaction*, *9*, 16–44.

Jonsson, A. (2013). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education*, *14*(1), 63–76.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86.

Kluger, A. N., & DeNisi, A. (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.

Klute, M., Apthorp, H., Harlacher, J., & Reale, M. (2017). *Formative assessment and elementary school student academic achievement: A review of the evidence*. Institute of Education Sciences, US Department of Education. https://ies.ed.gov/ncee/edlabs/regions/central/pdf/REL_2017259.pdf

Lee, H. W. (2011). The effects of generative learning strategy prompts and metacognitive feedback on learners' self-regulation, generation process, and achievement. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *71*(12-A), 1–180. https://etda.libraries.psu.edu/files/final_submissions/2268

Leton, E., Molanes-Lopez, E. M., Luque, M., & Conejo, R. (2018). Video podcast and illustrated text feedback in a web-based formative assessment environment. *Computer Applications in Engineering Education*, *26*(2), 187–202.

Lipnevich, A. A., & Smith, J. K. (2009). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied*, *15*(4), 319.

Lu, R., & Bol, L. (2007). A comparison of anonymous versus identifiable e-Peer review on college student writing performance and the extent of critical feedback. *Journal of Interactive Online Learning*, *6*(2), 100–115.

McDaniel, M. A., Bugg, J. M., Liu, Y. Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention?, *Journal of Experimental Psychology-Applied*, *21*(4), 370–382.

Mikheeva, M., Schneider, S., Beege, M., & Günter, D. R. (2019). Boundary conditions of the politeness effect in online mathematical learning. *Computers in Human Behavior*, *92*, 419–427.

Mitra, N. K., & Barua, A. (2015). Effect of online formative assessment on summative performance in integrated musculoskeletal system module. *BMC Medical Education*, *15*, 7.

Nicol, D. (2010). From monologue to dialogue: Improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education*, *35*(5), 501–517.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218.

OfS (2019a). *Student satisfaction rises but universities should do more to improve feedback*. https://www.offic eforstudents.org.uk/news-blog-and-events/press-and-media/student-satisfaction-rises-but-universities-should-do-more-to-improve-feedback/

OfS (2019b). *English higher education 2019: The Office for Students annual review*. https://www.officeforstuden ts.org.uk/annual-review-2019/a-high-quality-student-experience/

Panadero, E., & Alqassab, M. (2019). An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading. *Assessment & Evaluation in Higher Education*, *44*(8), 1253–1278. https://doi.org/10.1080/02602938.2019.1600186

Pennebaker, J., Gosling, S., & Ferrell, J. (2013). Daily online testing in large classes: Boosting college performance while reducing achievement gaps. *PLoS One*, *8*(11), e79774.

Perry, T., Davies, P., & Brady, J. (2020). Using video clubs to develop teachers' thinking and practice in oral feedback and dialogic teaching. *Cambridge Journal of Education*, *50*(5), 615–637.

Peterson, E., & Siadat, M. V. (2009). Combination of formative and summative assessment instruments in elementary algebra classes: A prescription for success. *Journal of Applied Research in the Community College*, *16*(2), 92–102.

Petrović, J., Pale, P., & Jeren, B. (2017). Online formative assessments in a digital signal processing course: Effects of feedback type and content difficulty on students learning achievements. *Education & Information Technologies*, *22*(6), 3047–3061.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, *28*(1), 4–13.

Rezaei, A. R. (2015). Frequent collaborative quiz taking and conceptual learning. *Active Learning in Higher Education*, *16*(3), 187–196.

Richards-Babb, M., Curtis, R., Ratcliff, B., Roy, A., & Mikalik, T. (2018). General chemistry student attitudes and success with use of online homework: traditional-responsive versus adaptive-responsive. *Journal of Chemical Education*, *95*(5), 691–699.

Roediger, H. L., & Karpicke, J. D. (2006). Test enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 77–84.

Sartain, A. F. (2018). *The frequency of testing and its effects on exam scores in a fundamental level baccalaureate nursing course*. (Ed.D.). The University of Alabama, Ann Arbor, MI.

Scalise, K., Douskey, M., & Stacy, A. (2018). Measuring learning gains and examining implications for student success in STEM. *Higher Education Pedagogies*, *3*(1), 183–195.

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, *143*(6), 565.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176–199.

Speckesser, S., Runge, J., Foliano, F., Bursnall, M., Hudson-Sharp, N., Rolfe, H., & Anders, J. (2018). *Embedding Formative Assessment: Evaluation report and executive summary*. London, UK: EEF.

Sweller, J., Ayres, J., & Kalyuga, S. (2011). *Cognitive load theory*. New York, NY: Springer-Verlag.

Torgerson, C. (2006). Publication bias: The Achilles heel of systematic reviews? *British Journal of Educational Studies*, *54*(1), 89–102.

van der Kleij, F. M., Eggen, T. J., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education*, *58*(1), 263–272.

Van der Schaaf, M., Baartman, L., Prins, F., Oosterbaan, A., & Schaap, H. (2013). Feedback dialogues that stimulate students' reflective thinking. *Scandinavian Journal of Educational Research*, *57*(3), 227–245.

Vogler, J. S., & Robinson, D. H. (2016). Team-based testing improves individual learning. *Journal of Experimental Education*, *84*(4), 787–803.

Vojdanoska, M., Cranney, J., & Newell, B. R. (2010). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology*, *24*(8), 1183–1195.

Weinstein, Y., & Sumeracki, M. (2018). *Understanding how we learn: A visual guide*. London, UK: Routledge.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14.

Wiliam, D. (2018). Feedback: At the heart of – but definitely not all of – formative assessment. In A. Lipnevic, & J. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 3–28). Cambridge, UK, Cambridge University Press.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 3087.

WWC (2020). *What works clearinghouse*. Retrieved from https://ies.ed.gov/ncee/wwc/. Accessed 18 September 2020.

Xiao, Y. (2011). *The effects of training in peer assessment on university students' writing performance and peer assessment quality in an online environment*. Retrieved from https://digitalcommons.odu.edu/teachinglearning_etds/44/. Accessed 18 September 2020.

Xiao, Y., & Lucking, R. (2008). The impact of two types of peer assessment on students' performance and satisfaction within a Wiki environment. *The Internet and Higher Education*, *11*(3–4), 186–193.

Zhang, X. (2018). *An examination of the effectiveness of peer feedback on Chinese university students' English writing performance*. (Ph.D.). Oakland University, Ann Arbor, MI.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education*, 9, e3292. https://doi.org/10.1002/rev3.3292