

Routledge
Taylor & Francis Group

Assessment & Evaluation in Higher Education

ISSN: 0260-2938 (Print) 1469-297X (Online) Journal homepage: www.tandfonline.com/journals/caeh20

From replacement to reform: why higher education should stop fighting exams and start making them formative

Ernesto Panadero

To cite this article: Ernesto Panadero (22 Sep 2026): From replacement to reform: why higher education should stop fighting exams and start making them formative, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2026.2734347

To link to this article: https://doi.org/10.1080/02602938.2026.2734347

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 22 Sep 2026.

Submit your article to this journal

Article views: 1033

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=caeh20

ASSESSMENT & EVALUATION IN HIGHER EDUCATION
https://doi.org/10.1080/02602938.2026.2734347

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

From replacement to reform: why higher education should stop fighting exams and start making them formative

Ernesto Panadero[a,b,c]

[a]Centre for Assessment Research Policy and Practice in Education (CARPE), School of Policy and Practice, Institute of Education, St. Patrick's Campus, Dublin City University, Dublin, Ireland; [b]Education Regulated Learning and Assessment (ERLA group), Facultad de Educación y Deportes, Universidad de Deusto, Bilbao, Spain; [c]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

ABSTRACT

For decades, influential voices within the formative assessment tradition have positioned exams as the primary obstacle to quality learning in higher education, and reform agendas from Bologna onwards have urged faculty to diversify away from them. This position paper argues that the strategy has failed: evidence from several countries shows that exam-centred assessment remains dominant and is not declining. Rather than continuing to advocate replacement, the paper proposes adding formative functions to the exams faculty already use. It identifies a conceptual error in the reform discourse, the conflation of assessment format with assessment purpose, and shows that the architects of formative assessment never intended to exclude exams from formative practice. The argument draws on the testing effect literature, where meta-analytic effect sizes range from 0.50 to 0.61. Five evidence-based practices are proposed: two-stage exams, exam wrappers, predictive self-assessment, cumulative design, and ungraded diagnostic midterms. A further section considers what generative artificial intelligence changes, and what it does not. The paper calls for a redirection of research, policy, and professional development effort.

KEYWORDS

Higher education assessment; exams; examinations; formative assessment; retrieval practice; assessment reform

Introduction

In a representative sample of Spanish public universities, 93.2% of course syllabi include at least one form of examination (Panadero et al. 2026a), and within courses followed over seven years the weight of final written exams in the final grade has increased (Panadero et al. 2026b). The persistence of exams is also documented in the United States (Lipnevich et al. 2020), the United Kingdom (Jessop and Maleckar 2016), and Finland (Halinen et al. 2014; Postareff et al. 2012), suggesting that this is not a local anomaly. Far from declining, the

CONTACT Ernesto Panadero ernesto.panadero@dcu.ie, ernesto.research@gmail.com Centre for Assessment Research, Policy and Practice in Education (CARPE), Dublin City University, Dublin, Ireland.
© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

exam-centred assessment profile is consolidating, absorbing courses that previously used more diversified designs (Panadero et al. 2026b). These findings emerge after decades of sustained scholarly and policy effort urging higher education to move beyond exams towards alternative assessment methods (Gibbs and Simpson 2005; Sambell, McDowell, and Brown 1997).

The assessment reform agenda has been remarkably consistent in its diagnosis: exams encourage surface learning, fail to develop higher-order competencies, generate anxiety, and reward memorisation over understanding (French, Dickerson, and Mulder 2024; Gibbs and Simpson 2005). The prescription has been equally consistent: replace exams with portfolios, projects, authentic tasks, and other methods that better align with desired learning outcomes. European policy reinforced this direction through the Bologna Process, which promoted competence-based assessment across participating countries, a recommendation that national legislation in countries such as Spain turned into a requirement. Yet the data suggest that this prescription, however well-intentioned, has not worked at scale.

This paper proposes that much of the higher education assessment reform discourse might have been asking the wrong question. Instead of 'how do we replace exams?' the question should be 'how do we make exams work for learning?'. The distinction may seem subtle, but its implications are substantial. The first question positions faculty who use exams as a problem to be solved. The second treats them as professionals working within real constraints who could benefit from evidence-based tools to enhance the formative potential of the assessment methods they already use.

This reframing rests on three observations. The first is empirical: exams have not declined despite decades of reform effort, and longitudinal and cross-national evidence shows exam-centred assessment to be a structural feature of higher education that policy has not changed. The second is conceptual: the reform discourse has conflated assessment format with assessment purpose, when the formative/summative distinction, as conceived by Scriven (1967) and elaborated by Bloom (1968), Black and Wiliam (2009), and Wiliam (2011), concerns how evidence is used, not which instrument generates it. The third draws on cognitive science: a large body of research on the testing effect shows that taking a test enhances learning and long-term retention, with meta-analytic effect sizes of 0.50 to 0.61 (Adesope, Trevisan, and Sundararajan 2017; Rowland 2014), a finding that the assessment field has largely ignored.

Two clarifications about the nature of this contribution are in order at the outset. First, this is a position paper. It reports no new empirical data. Second, what separates the present proposal from the wider formative assessment literature is its object. That literature has largely developed practices that sit alongside or in place of the graded exam. The claim advanced here is that the exam itself, graded or not, can be configured to carry formative function.

The persistence of exams

Exams remain central to higher education assessment across very different national systems. This is not an impression formed from local experience; it is a well-documented empirical reality.

Three studies of undergraduate course syllabi in Spanish public universities illustrate the pattern. Three examination modalities are distinguished throughout, following the coding framework used in these studies: the final written exam, sat at the end of the course; the partial exam, an intermediate or mid-term written test; and the practical exam, in which students perform a real-world task under examination conditions. Two cross-sectional waves, 1,693 syllabi across 48 universities in 2015-16 (Panadero et al. 2019) and 1,434 in 2022-23 (Panadero et al. 2026a), found final written exams in 70.2% and 72% of syllabi respectively, carrying the largest single weight in the final grade (about 57% to 59%), with partial exams in 36% to 39% of courses and practical exams in 15% to 18%, commonly added rather than substituted. When the three examination modalities are considered jointly, a calculation on the same dataset shows that 93.2% of syllabi (1,284 of the 1,378 for which assessment practices were fully coded, the analytic sample on which that study reports its individual percentages) include at least one form of examination, and only 6.8% prescribe none, confirming that exam-based assessment is near-universal in Spanish higher education. Self-assessment and peer-assessment remained marginal across both waves of data collection, appearing in fewer than 5% of courses. The third study (Panadero et al. 2026b) provides the first true longitudinal analysis by matching 1,347 syllabi across a seven-year span (2015-16 to 2022-23). Using latent transition analysis, it identifies distinct assessment profiles and tracks how courses move between them over time. Within these matched courses, the relative weight of the final written exam in the final grade rose significantly between the two waves, even as the cross-sectional average dipped slightly. An exam-centred profile not only persists but consolidates, absorbing courses that previously used diversified assessment designs. Where alternative assessment exists, it endures with 99% transition stability, yet it does not spread. Diversification, where it occurs, proceeds by adding components around the exam rather than by replacing it. The barrier is adoption, not sustainability, and the current strategy of exhorting faculty to abandon exams has not addressed it.

The pattern is not confined to Spain. In the United States, Lipnevich et al. (2020) analysed 250 syllabi across five academic disciplines and found that exams appeared in 78.8% of courses, with product criteria (of which exams are the most prevalent form) dominating grading in all disciplines except English and education. In the United Kingdom, where coursework predominates, Jessop and Maleckar (2016) analysed 18 degree programmes across eight universities and still found that examinations accounted for roughly 15% of assessment in humanities and professional programmes and 30.6% in science programmes, twice the proportion of the other discipline groups. In Finland, university life science teachers defaulted to summative written exams and often reused questions across years (Halinen et al. 2014; Postareff et al. 2012). The most recent Bologna Process Implementation Report acknowledges that assessment reform has progressed more slowly than structural reforms across the 49 countries of the European Higher Education Area (European Commission/EACEA/Eurydice 2024).

These findings suggest that exam dominance is not a peculiarity of any particular national system. It is a structural feature of how universities assess, across diverse institutional, cultural, and policy contexts. National regulatory frameworks tend to

reinforce this pattern rather than challenge it. In Spain, for example, the memoria that every degree programme must submit for verification has to specify, subject by subject, the evaluation systems to be used and the range of weight each may carry, and course syllabi must then conform to it (ANECA 2023). The requirement is more readily satisfied by formal, gradable formats than by open-ended formative practice. Similar regulatory pressures exist across the European Higher Education Area and beyond.

Why faculty (and students) value exams

The evidence just reviewed establishes that exam-centred assessment has not receded despite three decades of reform effort. Before asking why that reform has not succeeded, it is worth asking why exams have endured. The dominant assessment reform discourse tends to treat exam persistence as a problem to be diagnosed, an inertia to be overcome, or a failure of will on the part of faculty. This framing tells only part of the story, and there is empirical evidence that points the other way. Faculty choose exams because exams do certain things well, and students often agree with them.

When asked directly about their assessment practices, university teachers report valuing the comparability and standardisation that exams provide (Halinen et al. 2014; Postareff et al. 2012). All students are evaluated against the same criteria, at the same time, under the same conditions. In an institutional context where grades carry weight beyond the course itself, where students can formally contest them, where they affect scholarships, progression, and access to graduate study, this comparability is not a minor technicality. It is a core requirement of any defensible assessment. Brown (2004) found that teachers across school sectors hold conceptions of assessment that combine learning and accountability functions, and university faculty are no exception. Halinen et al. (2014) reported that life science teachers in Finland chose written exams in part because they perceived the format as treating all students equally, a justification framed in terms of equity rather than convenience.

There is also an efficiency argument that the formative assessment literature rarely engages with. As Broadbent, Panadero, and Boud (2018) note, university teachers operate under real time and workload constraints. A well-designed exam can assess sixty students in two hours and produce defensible grades in a week. An equivalent portfolio- or project-based system, with the feedback quality the formative tradition rightly demands, can take many times longer to mark. To present this as pedagogical conservatism is to misread the situation. It is closer to professional realism.

The pedagogical case for exams, however, goes beyond efficiency. Exams have genuine virtues as assessment instruments that the reform discourse has tended to overlook. They produce a clear, individual, time-bounded record of what each student can demonstrate without external assistance. They provide a definitive endpoint to a learning sequence, which has motivational and structural value: students study for exams in ways they often do not study for assignments without deadlines (Gibbs and Simpson 2005). And, as the cognitive science evidence reviewed below will show, the act of taking an exam is itself a learning event, with retrieval practice

producing some of the largest effect sizes documented in educational research (Adesope, Trevisan, and Sundararajan 2017; Roediger and Karpicke 2006). None of this is incompatible with a critical view of how exams are often used. But the instrument itself has more to recommend it than the dominant discourse has acknowledged.

Students themselves have a more nuanced relationship with exams than the anti-exam narrative often assumes. The evidence is thinner than it should be, and what exists is mixed. Struyven, Dochy, and Janssens (2005), reviewing student perceptions of assessment, found that alternative formats were generally seen as fairer than conventional exams in principle, while students voiced doubts about the reliability of self- and peer assessment even as they valued it; within conventional formats, the same review found that students preferred multiple-choice to essay examinations for their lower anxiety and higher expectancy of success. Sambell, McDowell, and Brown (1997) found that no format was perceived as fair by everyone. In practice, preferences tend to run towards the familiar. Van de Watering et al. (2008) found that even students enrolled in courses explicitly designed around innovative learning environments expressed a preference for traditional written examinations, valuing their predictability, clarity, and the perceived equity of being judged individually under controlled conditions. This is not a defense of exams against alternative formats. It is a reminder that students are not waiting to be liberated from exams. Many of them prefer them, for reasons that deserve to be taken seriously. It is also a reminder that what students themselves think about examinations, as opposed to what is assumed on their behalf, remains one of the least studied questions in this literature; the research agenda set out below returns to it.

Taken together, these perspectives complicate the standard reform narrative. Exams persist not only because of institutional inertia or insufficient training, but because they offer real value to the people who use them, both as instruments of evaluation and as learning experiences. Any serious proposal for assessment reform must begin from this recognition rather than against it.

Why the replacement narrative has failed

If three decades of reform discourse and explicit policy requirements have not reduced the dominance of exams, the question becomes: why? Three explanations, each reinforcing the others, might account for the gap between reform aspirations and assessment reality.

The first is a discourse-practice gap. Bologna changed the language of assessment without changing its substance. Syllabi adopted competence-based terminology and described assessment using the vocabulary of constructive alignment, but the underlying architecture remained exam-centred. Crucially, the gap is not attributable to faculty ignorance. Spanish university lecturers report assigning importance to monitoring student learning and feel reasonably capable of implementing assessment tasks for that purpose (Quesada-Serra, Rodríguez-Gómez, and Ibarra-Sáiz 2016). What they report not doing, and not feeling competent to do, is implementing participative assessment strategies such as self-assessment and peer-assessment. The gap is practical, not attitudinal.

This connects to a second explanation: faculty who use exams are not pedagogically naive. They are making choices that serve real institutional needs. The reasons set out in the previous section, comparability, defensibility, efficiency, and accountability, are not pretexts. Assessment choices are also shaped by disciplinary norms rather than by individual pedagogical preferences alone (Jessop and Maleckar 2016). A chemistry lecturer who uses exams is not unaware of alternative assessment; she is working within a disciplinary tradition where exams certify competence. Asking her to replace exams with portfolios is asking her to work against institutional structures, disciplinary traditions, student expectations, and her own professional judgement about what assessment is for, all at once. Any reform strategy that requires faculty to simultaneously change their assessment tool, their pedagogical identity, and their relationship to institutional accountability structures is unlikely to succeed at scale. The persistence and consolidation of exam-centred assessment documented earlier in this paper, despite sustained reform discourse since at least Boud (1990), confirms this pattern.

A related tension deserves brief acknowledgment, although a full treatment is reserved for future work. Much of the formative assessment literature proceeds as if grading were a secondary concern, something to be minimised or worked around. For the university teacher who must produce defensible numerical marks for sixty or three hundred students at the end of every semester, this framing is unrealistic. As Broadbent, Panadero, and Boud (2018, 308) observe, 'assessment is often constrained by the need for university teachers to produce numerical marks and grades to be formally recorded by the university by the end of the semester'. Grading is not a nuisance that good assessment design can eliminate; it is a legal, administrative, and pedagogical obligation that faculty feel acutely, and one which formative assessment advocates rarely engage with directly. Proposals that ignore this reality, however sound in the abstract, are unlikely to reach the practitioners they hope to influence. The approach advocated in this paper, by contrast, accepts the grading requirement as given and asks how the exam that produces the grade can simultaneously serve learning.

The third explanation is perhaps the most consequential, and the least discussed. The formative assessment tradition, from Black and Wiliam (1998) onward, has been extraordinarily productive: it has generated powerful theoretical frameworks, influenced policy worldwide, and improved practice in many classrooms. But it has systematically neglected exams as a vehicle for formative practice. The dominant framing positions exams as the antithesis of good assessment, something to be overcome rather than leveraged.

This amounts to a strategic error. If 93.2% of Spanish university syllabi include at least one form of examination and 78.8% of US syllabi rely on exams as a primary assessment tool (Lipnevich et al. 2020; Panadero et al. 2026a), and the formative assessment community has little to say about how to make those exams more formative, it has effectively conceded the territory where most assessment actually takes place. The field has spent decades reforming the narrow fraction of practice that does not rely on exams, while leaving the dominant modality it originally set out to transform essentially untouched, behind a narrative that sought to do away with what remains the most widely used assessment method in higher education.

The assessment literature routinely frames the relationship between exams and learning as adversarial. Harlen and Deakin Crick (2003), in a systematic review of school-level evidence, showed that high-stakes testing reduces intrinsic motivation and promotes surface learning; Gibbs and Simpson (2005) proposed conditions under which assessment supports learning that traditional exams rarely meet; and French, Dickerson, and Mulder (2024), reviewing high-stakes final examinations, concluded that the current reliance on them is poorly justified by the balance of evidence. These conclusions are sound for the designs they studied. But each examined exams as they are typically used, high-stakes, end-of-course, with minimal or delayed feedback and no follow-up. None examined exams redesigned with formative functions. The question this paper poses is therefore different: not whether poorly designed exams damage learning (they do), but whether exams redesigned with formative functions can support learning.

The irony is that the formative assessment field's own evidence points in this direction. Morris, Perry, and Wardle (2021), in the first international systematic review of formative assessment and feedback specifically in higher education, found that quizzing and testing constituted one of the most prevalent and promising areas of formative practice, with positive findings from both high-quality and lower-quality studies. They described quizzing as a 'good bet' for supporting student learning and attainment. Here, then, is a systematic review conducted within the formative assessment tradition, using formative assessment search terms and frameworks, arriving at the conclusion that testing works. The fact that this finding has not prompted a broader reconsideration of the field's stance towards exams is itself evidence of how deeply entrenched the anti-exam narrative has become.

The conceptual error: confusing instrument with purpose

At the heart of the exam-replacement narrative lies a conceptual error that has had outsized practical consequences: the conflation of assessment format with assessment purpose. The terms 'formative' and 'summative' were introduced by Scriven (1967) in the context of programme evaluation and adapted by Bloom (1969) for student assessment. From the outset, the distinction concerned purpose and use, not instrument type. Yet somewhere in the evolution of the field, 'formative' came to mean 'not an exam' and 'summative' became shorthand for 'exam'. This slippage has prevented the field from systematically investigating how exams might be redesigned to serve formative purposes.

The field's own leading theorists have resisted this conflation. Black and Wiliam (2009, 9) defined formative assessment as practice in which 'evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction'. Nothing in this definition excludes exams. If exam results are used to identify learning gaps, inform future teaching, and guide student self-regulation, the exam has served a formative function. Wiliam (2011) made the point even more explicitly with seven scenarios demonstrating that summative instruments, including standardised national tests, can serve formative purposes when their results are used to adapt instruction. Bennett (2011) offered what he called a 'more nuanced' view: well-designed summative tests support

learning through consolidation during preparation, retention enhancement through the testing effect, and identification of areas needing follow-up. Biggs (1998), in the same issue of *Assessment in Education*, went further still, arguing for integrating summative and formative assessment rather than treating them as dichotomous. The architect of constructive alignment saw no inherent contradiction between exams and learning.

The fact that Scriven, Bloom, Black, Wiliam, Bennett, and Biggs all saw a role for summative instruments in formative practice should have given the exam-replacement narrative pause. It did not.

There is a deeper precedent. Benjamin Bloom's mastery learning model (Bloom 1968) was, at its core, a formative use of exams. The strategy was straightforward: organise content into units, administer a brief test after each unit, diagnose mastery versus non-mastery, prescribe corrective activities, and retest. Bloom explicitly argued that formative tests should not be graded, because 'the use of grades on repeated progress tests prepares students for the acceptance of less than mastery' (Bloom 1968, 9). The problem, for Bloom, was never the test itself. It was the stakes attached to it and the absence of follow-up. When mastery learning is implemented with fidelity, its effect sizes are among the largest in education (Guskey 2007; Kulik, Kulik, and Bangert-Drowns 1990). What Bloom proposed nearly six decades ago, using tests as diagnostic tools for learning, remains largely unimplemented in higher education.

Perkins (2009) offered a complementary insight: learning is a consequence of thinking. What matters is whether the assessment provokes it, not its format: an exam requiring students to retrieve, apply, and evaluate knowledge provokes substantial cognitive work; a portfolio compiled by ticking criteria without reflection may not. Torrance (2007) documented this risk empirically: the very practices the formative assessment movement had championed as alternatives to traditional examination, explicit learning objectives, transparent criteria, and targeted feedback, paradoxically produced what he called 'criteria compliance' rather than genuine understanding. Format, then, is a poor proxy for learning quality. The relevant question is not 'exam or alternative?' but 'does this assessment make students think?' It is not the instrument that produces the learning; it is the use made of it.

The case from cognitive science: testing as learning

While influential voices in educational assessment have spent decades arguing against exams, a parallel field has accumulated overwhelming evidence that taking tests is one of the most effective strategies for promoting durable learning, by which cognitive scientists mean long-term retention and transfer (Karpicke 2012; Roediger and Butler 2011), not merely improved performance on the same test. The disconnect between these two bodies of research is remarkable, and closing it is essential for progress in assessment reform.

The testing effect refers to the finding that retrieving information from memory, as occurs during a test, strengthens retention more than restudying the same material for an equivalent period of time (Roediger and Karpicke 2006). In the seminal study,

students who took a free-recall test after studying prose passages retained substantially more at one week than students who restudied for the same time, even without feedback, and the students who restudied were the more confident of remembering the material.

The effect has been replicated across materials, populations, and settings. Two comprehensive meta-analyses quantify it. Rowland (2014) synthesised 159 effect sizes drawn from 61 studies published between 1975 and 2013, every one of which compared testing against a restudy control of equivalent duration, and reported a random-effects estimate of $g = 0.50$, 95% CI [0.42, 0.58], with substantial between-study heterogeneity and larger effects for recall than for recognition formats. Adesope, Trevisan, and Sundararajan (2017) synthesised 272 independent effect sizes from 118 experiments and reported $g = 0.61$ under a fixed-effects model and $g = 0.70$ under a random-effects model, with the effect moderated by question format (stronger for short-answer than multiple-choice), provision of feedback (stronger with feedback), and retention interval (larger effects at longer delays). Both estimates fall in the medium-to-large range. Dunlosky et al. (2013), evaluating ten learning techniques across materials, conditions, and criterion tasks, rated only two 'high utility': practice testing and distributed practice. Five techniques that students commonly use, including rereading, highlighting, and summarisation, received a 'low utility' rating. Assessment researchers tell faculty to reduce testing; cognitive scientists demonstrate that testing is among the most powerful learning tools available. Both fields are rigorous and evidence-based. They simply do not read each other's literature with sufficient attention.

A common objection is that the testing effect merely promotes rote recall, but the evidence contradicts this. Karpicke (2012) showed that retrieval practice produces meaningful, long-term learning: practicing retrieval changes the knowledge itself by improving its accessibility and flexibility. McDaniel et al. (2013) showed in middle-school science classes that quizzing improved performance on transfer questions as much as on identical ones, and that application-type quiz questions improved both definitional and application exam performance: testing promotes understanding, not just memorisation of answers.

The evidence from authentic classroom settings is also growing, although the university-specific base remains thin. Greving and Richter (2018) tested the effect during authentic university lectures and found that short-answer review questions, but not multiple-choice questions, produced a significant testing effect persisting up to 23 wk, supporting the recommendation that exams should be designed to require genuine retrieval rather than recognition. Roediger et al. (2011), working with sixth-grade social studies classes rather than university students, confirmed the pattern in a real classroom with actual course exams as dependent measures, showing that the benefits of quizzing persisted to end-of-semester exams administered one to two months later. These studies used real course materials, real students, and real assessments rather than the word lists and laboratory conditions of much of the earlier literature. The testing effect is a robust phenomenon with direct implications for how university exams should be designed, and one that research in universities has yet to test as thoroughly as it deserves.

Meeting faculty where they are: formative uses of exams

The evidence reviewed above converges on a simple conclusion: there are good reasons to keep exams and no realistic prospect of replacing them. The productive question is how to enhance their formative potential. What follows are five evidence-based practices that add formative functions to exams without requiring faculty to abandon the format. All five require minimal additional resources and are compatible with existing institutional structures, and they are not equally well evidenced. Table 1 summarises, for each, what changes, how it is implemented, how strong its evidence is, and the objection it most often meets. The practices are not mutually exclusive.

What distinguishes a formative exam from formative assessment more generally is neither the stakes attached to it nor whether it carries a grade. It is whether the assessment event is configured so that the evidence it produces is actually processed and acted upon. That configuration turns on a small number of things: what the student receives, what it sets in motion, whether they can act on it, when it arrives, and who takes part.

Measured against that, the conventional end-of-course exam fails on almost every count. What the student gets back is often no more than a number; where commented scripts are returned, they arrive when nothing further can be done with them, so what the exam sets in motion is a reaction to the grade rather than any processing of the content. It is terminal and high-stakes, and the only person involved in the judgement is the teacher. The exam is not formatively inert because it is an exam; it is formatively inert because it is configured in a way that prevents uptake. This is a question of assessment design rather than of assessment format, and each of the five practices changes one or more of those conditions. That is the reason they work.

In a two-stage collaborative exam, students first complete the exam individually, preserving the summative function, and then immediately work in small groups on the same or similar questions. The individual stage provides the grade; the group stage provides immediate peer feedback. Levy, Svoronos, and Klinger (2023) reported that the group stage score at Harvard University averaged approximately one standard deviation above the individual stage score. Gilley and Clarkston (2014) used a controlled crossover design and found that students showed significantly greater improvement on subsequent individual testing when they had taken the collaborative exam compared to individual-only testing, with equal benefits across upper, middle, and lower achievers. At the University of British Columbia, Wieman, Rieger, and Heiner (2014) described implementation in large introductory physics courses with 87% student approval.

Exam wrappers offer a different but complementary mechanism. These are structured reflection activities completed after students receive their graded exams, built around three questions: How did you prepare? What kinds of errors did you make? What will you change next time? (Lovett 2013). The concept originated with Achacoso's (2004) post-test analysis method, which included prediction of scores, Bloom's taxonomy analysis of errors, and identification of strategy-performance mismatches. Gezer-Templeton et al. (2017) implemented exam wrappers in a large

Table 1. Five formative uses of exams: what changes, how each is implemented, what the evidence shows, and the objection each meets.

Practice	What changes	Implementation	Evidence base	Usual objection and design response
Two-stage exam	Who takes part: peers join the judgement once the individual stage is over	Individual stage first, about two-thirds of the session; groups of three to five then answer a subset of the same questions; group stage weighted at up to 10% and unable to lower an individual's mark (Levy, Svoronos, and Klinger 2023; Wieman, Rieger, and Heiner 2014)	Evaluated in university courses; controlled crossover shows gains across ability levels (Gilley and Clarkston 2014)	Free-riding. The individual stage has already produced the grade; the group stage adds a bounded increment that cannot subtract.
Exam wrapper	What the returned exam sets in motion	One page returned with the marked exam (how did you prepare, what errors, what will change); collected, then handed back when preparation for the next exam begins (Lovett 2013)	Evaluated in university courses; gains concentrated in the middle of the grade distribution (Gezer-Templeton et al. 2017); no effect when used once without follow-through (Soicher and Gurung 2017)	Superficial completion. Works only if the wrapper is revisited before the next exam.
Predictive self-assessment	The student's own role, and what comes back: content-level information rather than a single mark	Short form before the exam is returned: predicted grade and a confidence rating for each topic; returned with the marked exam so that prediction and result sit side by side (Achacoso 2004)	Components well documented in the self-assessment and calibration literature; the package with exams rarely evaluated as such	Overestimators are the least likely to self-correct. The gap must be put in front of them, not left for them to notice.
Cumulative design	Timing and sequence, not content	Each exam after the first reserves a quarter to a third of its questions for earlier material; the final draws on the whole course	Mechanism well established, including in school classrooms (Roediger et al. 2011); cumulative university exams as a deliberate design rarely evaluated	Coverage. A modest loss of measurement on recent material for a substantial gain in retention of earlier material.
Ungraded diagnostic midterm	Stakes removed; a diagnosis comes back rather than a grade	Sat under normal conditions, returned with performance by topic, followed by targeted correctives and, where possible, a brief retest (Bloom 1968)	Mastery-learning evidence largely from school and early college settings (Guskey 2007; Kulik, Kulik, and Bangert-Drowns 1990); ungraded diagnostic midterms in universities rarely evaluated	Students will not prepare for what does not count. Draw part of the graded final from the same question pool.

introductory course and found that students improved study strategies across the semester, with particular benefits for students in the middle of the grade distribution. Exam wrappers turn the exam from an endpoint into a learning cycle, requiring minimal class time and bridging exam practice to self-regulated learning.

The step that matters most is the one most often omitted: the completed wrappers must come back to students when they begin preparing for the next exam (Lovett 2013). Soicher and Gurung (2017) found no effect on scores, grades, or measured metacognition when wrappers were used once without follow-through, and Lovett (2013) reported gains only among students who used them in more than one course. The practice works when the reflection is revisited, not as a form filled in once and forgotten.

A more nuanced intervention is predictive self-assessment. In its simplest form, students predict their score before the exam is returned, which activates metacognitive monitoring and reveals calibration accuracy, the degree to which their confidence matches their performance; students who overestimate frequently suffer from what Glenberg, Wilkinson, and Epstein (1982) termed the 'illusion of knowing'. The practice is more powerful when extended from the global grade to the content: students identify, before the exam, which topics they feel confident about and which they remain uncertain on. This combination of self-grading and content-level self-assessment connects directly to evaluative judgement, the capacity to make decisions about the quality of one's own and others' work (Tai et al. 2018), and to the broader literature on self-assessment as a formative practice (Andrade 2019). Every exam then returns objective, criterion-referenced feedback on the accuracy of both the grade prediction and the content-level self-evaluation.

Cumulative and spaced exam design offers yet another avenue. Rather than testing each unit separately and moving on, cumulative design requires students to revisit previously assessed material, exploiting two additional high-utility learning strategies: distributed practice and interleaving (Dunlosky et al. 2013). The mechanism is well established: Roediger et al. (2011) showed, in school classrooms, that material retrieved repeatedly across a semester was retained better at delayed exams than material studied but not retrieved. Evidence on cumulative university exams as a deliberate design is sparser than evidence on the mechanism they exploit. Structuring a course so that each assessment builds cumulatively on previous content transforms the assessment sequence from isolated judgements into a distributed retrieval practice schedule.

Finally, ungraded diagnostic midterms return to Bloom's (1968) original proposal: administer the test, diagnose mastery, prescribe correctives, but do not grade it. Removing the grade eliminates the motivational damage of high-stakes testing documented in schools by Harlen and Deakin Crick (2003) while preserving the retrieval benefit. The test becomes a checkpoint rather than a judgement. It suits large courses where individual feedback is impractical: the exam itself provides the feedback by showing each student what they know and what they do not. The mastery learning evidence (Guskey 2007; Kulik, Kulik, and Bangert-Drowns 1990) comes largely from school and early college settings; direct evaluation of ungraded diagnostic midterms in university courses is rare.

To see how the practices combine, consider a one-semester course with two midterms and a final. Before each exam, students complete a one-page form predicting

their grade and rating their confidence topic by topic. The first midterm is diagnostic and ungraded; it is returned with a topic-level breakdown, and its wrapper asks students to set their predictions against their results and to record what they will change. The second midterm is graded and cumulative, drawing a third of its questions from material the first covered. The final is cumulative and two-stage, with the group stage weighted at ten percent. Everything about how the exams are used has changed. The exam format has not changed at all. That is the point.

None of these practices, however, will reach the faculty who need them without systematic professional development. The assessment reform movement has invested heavily in training faculty to adopt alternative assessment methods, and the results have been modest. But there has been almost no investment in training faculty to use their existing exams more formatively. The scarcity is documented: a nationwide analysis of professional development courses on assessment across all 50 Spanish public universities found only 82 courses available, with self-assessment, peer-assessment, and feedback systematically underrepresented in the offer (Fernández Ruiz and Panadero 2023). A count of course titles in that dataset makes the point: five of the 82 courses mention examinations, four of them about delivering exams online and one about generating questions from templates, and none about formative uses, against 21 courses on other assessment methods, 17 on rubrics, and three on portfolios. Titles and descriptors were all that was available, so a generalist course may treat examinations without saying so, but the asymmetry is unmistakable. A workshop titled 'Making your exams work harder for learning' would likely draw far more faculty than one titled 'Replacing your exams with authentic assessment', because it speaks to what faculty actually do. These tools will remain in the pages of journals unless institutions create the conditions for faculty to learn them.

This paper does not argue that exams are sufficient for all assessment purposes; clinical skills, laboratory technique, collaborative problem-solving, and sustained independent research all require other instruments. The argument is narrower: where exams are already used, as they are in most university courses, they should be used better. The proposal is not 'more exams' but 'better exams'.

Exams in the age of generative AI

Generative artificial intelligence has disrupted assessment practices across higher education. Take-home assignments, essays, reports, and other open-ended tasks can be substantially completed or enhanced by AI tools. Institutions have responded with a range of strategies, including redesigning assessments for 'authenticity', on the assumption that authentic, context-dependent tasks are more resistant to AI-assisted completion than traditional exams (Francis, Jones, and Smith 2025).

Recent evidence undermines this assumption. Kofinas, Tsay, and Pike (2025), in experiments at two UK universities, found that markers were generally unable to distinguish assessments with GenAI input from those without, and that the authenticity of an assessment made no difference to the ability to detect or safeguard against GenAI use. Francis, Jones, and Smith (2025), reviewing the broader landscape, reached a similar conclusion: GenAI poses challenges to assessment integrity that no single format can solve.

It is tempting to conclude from this that in-person, proctored exams remain one of the few contexts in which individual, unaided cognitive work can still be verified. That conclusion is no longer safe. Corbin, Sharpe, and Dawson (2026) document that AI-enabled smart glasses, now mass-market products indistinguishable from ordinary eyewear, can display generated text in the wearer's line of sight, process speech, and read examination materials, without producing any reliable external signal. Physical exclusion, the logic underpinning the current return to exams and oral assessment, works only if AI use can be cleanly separated from the student's own activity and leaves some trace when it occurs. Wearable AI weakens both assumptions. This eliminates the argument that replacing exams with alternative assessment solves the integrity challenges posed by GenAI. Both formats are vulnerable, and neither is inherently more secure. The implication is not that exams should be abandoned. It is that the case for exams cannot be a custodial one. If neither open assessment nor proctored examination can be 'secured' against AI assistance, then security, which Dawson (2021) has argued is in any case a matter of degree rather than a property an assessment either has or lacks, is the wrong ground on which to defend either. What remains is the pedagogical ground, which is the argument of this paper: exams are worth keeping because a well-designed exam produces learning, not because it can be policed.

The question, then, is how to design an assessment system in which exams and alternatives each have a role. In-person exams, redesigned with formative functions, remain the setting in which students are most likely to engage in the unaided retrieval and reasoning that retains pedagogical value (Perkins 2009), even where such engagement can no longer be guaranteed. Corbin, Sharpe, and Dawson (2026) further note that wearable AI has legitimate uses, from translation to hearing support, which makes blanket prohibition both impractical and educationally questionable. Other assessment modalities can and should evaluate students' ability to work competently with AI tools, a competence graduates will need across most professional domains. Treating AI primarily as a threat to be policed misses the point: the issue is not cheating but the validity of what assessments actually measure (Dawson et al. 2024). Two-stage exams, exam wrappers, and predictive self-assessment are particularly well suited to the first kind of space, since what they add, reflection on one's own errors and calibration of one's own confidence, cannot be outsourced without becoming meaningless.

Implications and future directions

The argument presented in this paper carries implications for policy, practice, and research in higher education assessment. First, considering policy: assessment reform should stop framing improvement as 'moving away from exams' and instead frame it as 'making all assessment, including exams, work for learning'. Quality assurance frameworks could track indicators of formative exam design alongside the diversity of assessment methods they currently count. Investment in professional development should include training on formative exam design.

Second, considering practice: the practices can be adopted incrementally, from the one-page wrapper to the two-stage final, and they are additive rather than

substitutive. The message is not 'you are doing it wrong' but 'here are evidence-based ways to do more with what you already do'.

Third, considering research: several lines of inquiry would strengthen the evidence base. To begin with, we need randomised controlled trials of formative exam interventions in university settings, comparing courses that use exam wrappers, two-stage formats, or cumulative design against matched courses that use the same exams without these additions. Most testing effect research is laboratory-based: Rowland (2014) excluded classroom studies from his meta-analysis, and only around 11% of the experiments Adesope, Trevisan, and Sundararajan (2017) coded took place in classrooms; the university evidence remains largely observational or quasi-experimental. Second, longitudinal tracking across a degree programme would establish whether the benefits observed in semester-long studies persist and accumulate. Third, the student perspective on examinations is underresearched: we know more about what reformers assume students want from assessment than about what students report wanting, and the few studies that ask them (Jessop and Maleckar 2016; Struyven, Dochy, and Janssens 2005; Van de Watering et al. 2008) point in directions the reform discourse does not anticipate. Finally, the professional development question deserves its own research program: what training models are most effective for helping faculty adopt formative exam practices, and what institutional conditions support or hinder their uptake?

Conclusion

For three decades, influential voices within the formative assessment tradition have presented exams to higher education faculty as a problem to be solved by replacement. Faculty have, for the most part, continued to use them. The evidence reviewed in this paper suggests that this was not unreasonable: exams serve real institutional needs, and the alternatives proposed by the field have not diffused at scale despite sustained policy support.

But faculty were not entirely right either. Using exams as isolated, end-of-course, high-stakes events with minimal feedback is a waste of one of the most powerful learning tools available. The question is not whether to use exams, but how to use them well.

This paper has argued for a paradigm shift: from replacement to reform. Instead of fighting a battle against exams that we have been losing for three decades, we should invest our energy in making exams serve learning. The tools exist. They are evidence-based, scalable, and compatible with existing institutional structures. What is needed is not more exhortation to abandon exams but more support for faculty who want to use them better.

The exam is not going away. It is time we made it work.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Spanish Ministry of Science, Innovation and Universities under grant PID2024-156609NB-I00.

ORCID

Ernesto Panadero http://orcid.org/0000-0003-0859-3616

Data availability statement

No new data were generated for this article.

References

Achacoso, M. V. 2004. "Post-Test Analysis: A Tool for Developing Students' Metacognitive Awareness and Self-Regulation." *New Directions for Teaching and Learning* 2004 (100): 115–119. https://doi.org/10.1002/tl.179.

Adesope, O. O., D. A. Trevisan, and N. Sundararajan. 2017. "Rethinking the Use of Tests: A Meta-Analysis of Practice Testing." *Review of Educational Research* 87 (3): 659–701. https://doi.org/10.3102/0034654316689306.

Andrade, H. L. 2019. "A Critical Review of Research on Student Self-Assessment." *Frontiers in Education* 4: 87. https://doi.org/10.3389/feduc.2019.00087.

ANECA. 2023. Guía de apoyo para la elaboración de la memoria para la solicitud de verificación de títulos oficiales(Grado y Máster). Madrid: Agencia Nacional de Evaluación de la Calidad y Acreditación.

Bennett, R. E. 2011. "Formative Assessment: A Critical Review." *Assessment in Education: Principles, Policy & Practice* 18 (1): 5–25. https://doi.org/10.1080/0969594X.2010.513678.

Biggs, J. 1998. "Assessment and Classroom Learning: A Role for Summative Assessment?" *Assessment in Education: Principles, Policy & Practice* 5 (1): 103–110. https://doi.org/10.1080/0969595980050106.

Black, P., and D. Wiliam. 1998. "Assessment and Classroom Learning." *Assessment in Education: Principles, Policy & Practice* 5 (1): 7–74. https://doi.org/10.1080/0969595980050102.

Black, P., and D. Wiliam. 2009. "Developing the Theory of Formative Assessment." *Educational Assessment, Evaluation and Accountability* 21 (1): 5–31. https://doi.org/10.1007/s11092-008-9068-5.

Bloom, B. S. 1968. "Learning for Mastery." *Evaluation Comment* 1 (2): 1–12.

Bloom, B. S. 1969. "Some Theoretical Issues Relating to Educational Evaluation." In *Educational Evaluation: New Roles, New Means*, edited by R. W. Tyler, 26–50. Chicago, IL: University of Chicago Press. https://doi.org/10.1177/016146816907001003.

Boud, D. 1990. "Assessment and the Promotion of Academic Values." *Studies in Higher Education* 15 (1): 101–111. https://doi.org/10.1080/03075079012331377621.

Broadbent, J., E. Panadero, and D. Boud. 2018. "Implementing Summative Assessment with a Formative Flavour: A Case Study in a Large Class." *Assessment & Evaluation in Higher Education* 43 (2): 307–322. https://doi.org/10.1080/02602938.2017.1343455.

Brown, G. T. L. 2004. "Teachers' Conceptions of Assessment: Implications for Policy and Professional Development." *Assessment in Education: Principles." Policy & Practice* 11 (3): 301–318. https://doi.org/10.1080/0969594042000304609.

Corbin, T., S. Sharpe, and P. Dawson. 2026. "On AI Glasses and Wearable AI in Assessment." *Assessment & Evaluation in Higher Education*: 1–17. Advance online publication. https://doi.org/10.1080/02602938.2026.2661367.

Dawson, P. 2021. *Defending Assessment Security in a Digital World: Preventing e-Cheating and Supporting Academic Integrity in Higher Education*. Abingdon: Routledge.

Dawson, P., M. Bearman, M. Dollinger, and D. Boud. 2024. "Validity Matters More than Cheating." *Assessment & Evaluation in Higher Education* 49 (7): 1005–1016. https://doi.org/10.1080/02602938.2024.2386662.

Dunlosky, J., K. A. Rawson, E. J. Marsh, M. J. Nathan, and D. T. Willingham. 2013. "Improving Students' Learning with Effective Learning Techniques: Promising Directions from Cognitive and Educational Psychology." *Psychological Science in the Public Interest: A Journal of the American Psychological Society* 14 (1): 4–58. https://doi.org/10.1177/1529100612453266.

European Commission/EACEA/Eurydice. 2024. *The European Higher Education Area in 2024: Bologna Process Implementation Report*. Luxembourg: Publications Office of the European Union.

Fernández Ruiz, J., and E. Panadero. 2023. "Assessment Professional Development Courses for University Teachers: A Nationwide Analysis Exploring Length, Evaluation and Content." *Assessment & Evaluation in Higher Education* 48 (4): 485–501. https://doi.org/10.1080/02602938.2022.2099811.

Francis, N. J., S. Jones, and D. P. Smith. 2025. "Generative AI in Higher Education: Balancing Innovation and Integrity." *British Journal of Biomedical Science* 81: 14048. https://doi.org/10.3389/bjbs.2024.14048.

French, S., A. Dickerson, and R. A. Mulder. 2024. "A Review of the Benefits and Drawbacks of High-Stakes Final Examinations in Higher Education." *Higher Education* 88 (3): 893–918. https://doi.org/10.1007/s10734-023-01148-z.

Gezer-Templeton, P. G., E. J. Mayhew, D. S. Korte, and S. J. Schmidt. 2017. "Use of Exam Wrappers to Enhance Students' Metacognitive Skills in a Large Introductory Food Science and Human Nutrition Course." *Journal of Food Science Education* 16 (1): 28–36. https://doi.org/10.1111/1541-4329.12103.

Gibbs, G., and C. Simpson. 2005. "Conditions Under Which Assessment Supports Students' Learning." *Learning and Teaching in Higher Education* 1: 3–31.

Gilley, B. H., and B. Clarkston. 2014. "Collaborative Testing: Evidence of Learning in a Controlled In-Class Study of Undergraduate Students." *Journal of College Science Teaching* 43 (3): 83–91.

Glenberg, A. M., A. C. Wilkinson, and W. Epstein. 1982. "The Illusion of Knowing: Failure in the Self-Assessment of Comprehension." *Memory & Cognition* 10 (6): 597–602. https://doi.org/10.3758/BF03202442.

Greving, S., and T. Richter. 2018. "Examining the Testing Effect in University Teaching: Retrievability and Question Format Matter." *Frontiers in Psychology* 9: 2412. https://doi.org/10.3389/fpsyg.2018.02412.

Guskey, T. R. 2007. "Closing Achievement Gaps: Revisiting Benjamin S. Bloom's 'Learning for Mastery." *Journal of Advanced Academics* 19 (1): 8–31. https://doi.org/10.4219/jaa-2007-704.

Halinen, K., M. Ruohoniemi, N. Katajavuori, and V. Virtanen. 2014. "Life Science Teachers' Discourse on Assessment: A Valuable Insight into the Variable Conceptions of Assessment in Higher Education." *Journal of Biological Education* 48 (1): 16–22. https://doi.org/10.1080/00219266.2013.799082.

Harlen, W., and R. Deakin Crick. 2003. "Testing and Motivation for Learning." *Assessment in Education: Principles, Policy & Practice* 10 (2): 169–207. https://doi.org/10.1080/0969594032000121270.

Jessop, T., and B. Maleckar. 2016. "The Influence of Disciplinary Assessment Patterns on Student Learning: A Comparative Study." *Studies in Higher Education* 41 (4): 696–711. https://doi.org/10.1080/03075079.2014.943170.

Karpicke, J. D. 2012. "Retrieval-Based Learning: Active Retrieval Promotes Meaningful Learning." *Current Directions in Psychological Science* 21 (3): 157–163. https://doi.org/10.1177/0963721412443552.

Kofinas, A. K., C. H.-H. Tsay, and D. Pike. 2025. "The Impact of Generative AI on Academic Integrity of Authentic Assessments within a Higher Education Context." *British Journal of Educational Technology* 56 (6): 2522–2549. https://doi.org/10.1111/bjet.13585.

Kulik, C. C., J. A. Kulik, and R. L. Bangert-Drowns. 1990. "Effectiveness of Mastery Learning Programs: A Meta-Analysis." *Review of Educational Research* 60 (2): 265–299. https://doi.org/10.3102/00346543060002265.

Levy, D., T. Svoronos, and M. Klinger. 2023. "Two-Stage Examinations: Can Examinations Be More Formative Experiences?" *Active Learning in Higher Education* 24 (2): 79–94. https://doi.org/10.1177/1469787418801668.

Lipnevich, A. A., T. R. Guskey, D. M. Murano, and J. K. Smith. 2020. "What Do Grades Mean? Variation in Grading Criteria in American College and University Courses." *Assessment in Education: Principles, Policy & Practice* 27 (5): 480–500. https://doi.org/10.1080/0969594X.2020.1799190.

Lovett, M. C. 2013. "Make Exams Worth More than the Grade: Using Exam Wrappers to Promote Metacognition." In *Using Reflection and Metacognition to Improve Student Learning*, edited by M. Kaplan, N. Silver, D. LaVague-Manty, and D. Meizlish, 18–52. Sterling, VA: Stylus.

McDaniel, M. A., R. C. Thomas, P. K. Agarwal, K. B. McDermott, and H. L. Roediger. 2013. "Quizzing in Middle-School Science: Successful Transfer Performance on Classroom Exams." *Applied Cognitive Psychology* 27 (3): 360–372. https://doi.org/10.1002/acp.2914.

Morris, R., T. Perry, and L. Wardle. 2021. "Formative Assessment and Feedback for Learning in Higher Education: A Systematic Review." *Review of Education* 9 (3): E 3292. https://doi.org/10.1002/rev3.3292.

Panadero, E., A. Amezua-Urrutia, J. Fernández Ruiz, C. Rodríguez-Hernández, and J. Fraile. 2026a. "Assessment Practices in Spanish Universities: A Nationwide Update Seven Years On." *Assessment & Evaluation in Higher Education* : 1–21. Advance online publication. https://doi.org/10.1080/02602938.2026.2621276.

Panadero, E., A. Amezua-Urrutia, J. Fernández Ruiz, C. Rodríguez-Hernández, and J. Fraile. 2026b. "Spanish University Assessment Practices over Time: A Nationwide Longitudinal Syllabus Analysis (2015–2016 Vs. 2022–2023)." *Assessment & Evaluation in Higher Education*: 1–20. Advance online publication. https://doi.org/10.1080/02602938.2026.2707269.

Panadero, E., J. Fraile, J. Fernández Ruiz, D. Castilla-Estévez, and M. A. Ruiz. 2019. "Spanish University Assessment Practices: Examination Tradition with Diversity by Faculty." *Assessment & Evaluation in Higher Education* 44 (3): 379–397. https://doi.org/10.1080/02602938.2018.1512553.

Perkins, D. 2009. *Making Learning Whole: How Seven Principles of Teaching Can Transform Education*. San Francisco, CA: Jossey-Bass.

Postareff, L., V. Virtanen, N. Katajavuori, and S. Lindblom-Ylänne. 2012. "Academics' Conceptions of Assessment and Their Assessment Practices." *Studies in Educational Evaluation* 38 (3–4): 84–92. https://doi.org/10.1016/j.stueduc.2012.06.003.

Quesada-Serra, V., G. Rodríguez-Gómez, and M. S. Ibarra-Sáiz. 2016. "What Are We Missing? Spanish Lecturers' Perceptions of Their Assessment Practices." *Innovations in Education and Teaching International* 53 (1): 48–59. https://doi.org/10.1080/14703297.2014.930353.

Roediger, H. L., III, and A. C. Butler. 2011. "The Critical Role of Retrieval Practice in Long-Term Retention." *Trends in Cognitive Sciences* 15 (1): 20–27. https://doi.org/10.1016/j.tics.2010.09.003.

Roediger, H. L., III, and J. D. Karpicke. 2006. "Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention." *Psychological Science* 17 (3): 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x.

Roediger, H. L., III, P. K. Agarwal, M. A. McDaniel, and K. B. McDermott. 2011. "Test-Enhanced Learning in the Classroom: Long-Term Improvements from Quizzing." *Journal of Experimental Psychology. Applied* 17 (4): 382–395. https://doi.org/10.1037/a0026252.

Rowland, C. A. 2014. "The Effect of Testing versus Restudy on Retention: A Meta-Analytic Review of the Testing Effect." *Psychological Bulletin* 140 (6): 1432–1463. https://doi.org/10.1037/a0037559.

Sambell, K., L. McDowell, and S. Brown. 1997. "'But Is It Fair?': An Exploratory Study of Student Perceptions of the Consequential Validity of Assessment." *Studies in Educational Evaluation* 23 (4): 349–371. https://doi.org/10.1016/S0191-491X(97)86215-3.

Scriven, M. 1967. "The Methodology of Evaluation." In *Perspectives of Curriculum Evaluation*, edited by R. W. Tyler, R. M. Gagné, and M. Scriven, 39–83. Chicago, IL: Rand McNally.

Soicher, R. N., and R. A. R. Gurung. 2017. "Do Exam Wrappers Increase Metacognition and Performance? A Single Course Intervention." *Psychology Learning & Teaching* 16 (1): 64–73. https://doi.org/10.1177/1475725716661872.

Struyven, K., F. Dochy, and S. Janssens. 2005. "Students' Perceptions about Evaluation and Assessment in Higher Education: A Review." *Assessment & Evaluation in Higher Education* 30 (4): 325–341. https://doi.org/10.1080/02602930500099102.

Tai, J., R. Ajjawi, D. Boud, P. Dawson, and E. Panadero. 2018. "Developing Evaluative Judgement: Enabling Students to Make Decisions about the Quality of Work." *Higher Education* 76 (3): 467–481. https://doi.org/10.1007/s10734-017-0220-3.

Torrance, H. 2007. "Assessment as Learning? How the Use of Explicit Learning Objectives, Assessment Criteria and Feedback in Post-Secondary Education and Training Can Come to Dominate Learning." *Assessment in Education: Principles, Policy & Practice* 14 (3): 281–294. https://doi.org/10.1080/09695940701591867.

Van de Watering, G., D. Gijbels, F. Dochy, and J. Van der Rijt. 2008. "Students' Assessment Preferences, Perceptions of Assessment and Their Relationships to Study Results." *Higher Education* 56 (6): 645–658. https://doi.org/10.1007/s10734-008-9116-6.

Wieman, C., G. Rieger, and C. Heiner. 2014. "Physics Exams That Promote Collaborative Learning." *The Physics Teacher* 52 (1): 51–53. https://doi.org/10.1119/1.4849159.

Wiliam, D. 2011. "What Is Assessment for Learning?" *Studies in Educational Evaluation* 37 (1): 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001.