Higher Education
https://doi.org/10.1007/s10734-023-01148-z

A review of the benefits and drawbacks of high-stakes final examinations in higher education

Sarah French[1] · Ashton Dickerson[2] · Raoul A. Mulder[1,2]

Accepted: 20 November 2023

Abstract
High-stakes examinations enjoy widespread use as summative assessments in higher education. We review the arguments for and against their use, across seven common themes: memory recall and knowledge retention; student motivation and learning; authenticity and real-world relevance; validity and reliability; academic misconduct and contract cheating; stress, anxiety and wellbeing; and fairness and equity. For each theme, we evaluate empirical evidence for the perceived pedagogical benefits and pedagogical drawbacks of high-stakes examinations. We find that relatively few of the perceived academic benefits of high-stakes examinations have a strong evidence base. Support for their use is largely rooted in opinion and pragmatism, rather than being justified by scientific evidence or pedagogical merit. By contrast, there is substantial evidence for pedagogical drawbacks of high-stakes summative examinations. We conclude that the current heavy reliance on high-stakes final examinations in many university subjects is poorly justified by the balance of empirical evidence.

Keywords High-stakes examinations · Assessment · Learning and teaching · Higher education

Introduction

Assessment plays a crucial role in higher education, providing a means to certify levels of achievement, verify learning outcomes, and test student knowledge. However, assessment has considerable additional educational value due to its potential to motivate, facilitate and enhance learning (Carless et al., 2017; Entwistle & Entwistle, 1991; Kickert et al., 2022; Marton & Säljö, 1997; Ramsden, 1997), and lay the foundations for future learning

✉ Raoul A. Mulder
r.mulder@unimelb.edu.au

Sarah French
frenchs@unimelb.ed.au

Ashton Dickerson
ashton.l.dickerson@gmail.com

[1] Centre for the Study of Higher Education, Faculty of Education, University of Melbourne, Victoria 3010, Australia

[2] School of BioSciences, Faculty of Science, University of Melbourne, Victoria 3010, Australia

Published online: 01 December 2023

(Boud & Falchikov, 2006; Boud, 1995, 2000). The way in which students are assessed also has profound implications for both student wellbeing (Baik et al., 2019; Jones et al., 2021; Slavin et al., 2014) and student engagement (Vaughan, 2014) and, arguably more than any other aspect of teaching, signals to students what is valued by their teachers, the discipline, and the institution.

The challenge of designing effective assessment is a perennial problem for universities, and one of the primary issues of concern identified by higher education quality assurance bodies. In reviews conducted by the Quality Assurance Agency for Higher Education in the UK, for example, deficiencies related to assessment practices have persistently emerged as the main criticism of university courses, especially the 'very narrow range of assessment methods in use and over-reliance on traditional examinations' (Boud & Falchikov, 2006, p. 402). This over-reliance on examinations is problematic for two key reasons. Firstly, it constrains diversity in assessment methods. Such diversity is necessary to assess a broad range of learning outcomes, provide a multi-dimensional understanding of student's skills and knowledge, maintain student engagement, and involve students in learning activities that lead to higher order thinking and a deeper understanding of content (Biggs et al., 2022). Secondly, high-stakes final examinations tend to serve a purely summative function, which becomes an issue when they dominate the curriculum at the expense of opportunities for formative assessment and feedback. It is important that formative assessment and feedback feature prominently in curriculum design (Morris et al., 2021) to allow students to advance their learning by actively engaging with and implementing feedback (Henderson et al., 2020; Winstone & Carless, 2020).

In the scholarly literature on assessment in higher education, questions relating to the pedagogical value of final examinations surface repeatedly. For example, John Biggs (2001, p. 234) argues that "invigilated examinations are hard to justify educationally" citing concerns about plagiarism and contract cheating as the leading "distorted priority" for their ongoing use. Scholars such as Gibbs (1992) and Ramsden (1992) warn against the reliability of examinations as a measurement of student learning, noting that questions assessing the recall of facts can often be answered without an understanding of the fundamental principles of the topic or a more complex understanding of the ways in which concepts are integrated in real-world scenarios. Such critiques are by no means new. Indeed, examinations have received criticism since their inception in Imperial China when they were widely criticised for their emphasis on rote memorisation, testing of skills rather than knowledge, the prevalence of cheating, and for cases of mental disorders that were anecdotally attributed to failing the high-pressured examinations (Kellaghan & Greaney, 2019). Similar criticisms were made of the written examinations introduced at Oxford and Cambridge in the nineteenth century which were perceived to dissuade originality through their focus on recall and to contribute to social stratification by benefiting the most privileged (Kellaghan & Greaney, 2019). Yet examinations continued to hold a privileged place in universities, and in fact, a global feature of educational systems was the increase in prevalence of examinations throughout the nineteenth and twentieth centuries (Kellaghan & Greaney, 2019), despite strong criticisms of their low reliability (Hartog & Rhodes, 1936).

Since the 1970s, there has been a shift away from the use of high-stakes final examinations in many countries, including New Zealand (Bassey, 1971), Finland (Sahlberg & Hargreaves, 2011), Australia, and the UK (Richardson, 2015a). This shift is in part a response to the growing concern that heavily weighted summative examinations may negatively impact student learning and wellbeing (Ecclestone, 1999; Jones et al., 2021; Pascoe et al., 2020) and debate about their efficacy and validity as assessment instruments (Knight, 2002). It is also related to broader macro-level processes, including the internationalization of higher education which brings global dimensions into the curriculum that impact assessment design (Jamil et al., 2021), and digitalization which has opened-up new possibilities for more diverse and creative assessment methods that can be employed at scale. However, in many systems, high-stakes examinations remain strongly entrenched. For example, Wong et al. (2020) observe that in Singapore attempts to introduce new modes of assessment have been constrained by a high-stakes examination culture that has been a key feature of the national education system since the 1960s. Similarly, researchers in China (Chen et al., 2020; Song, 2016; Wang et al., 2022; Wang & Brown, 2014), Korea (Kwon et al., 2017), South Africa (Mutereko, 2018), the U.S (Berliner, 2011; Fook & Sidhu, 2014; Gorgodze & Chakhaia, 2021), and Canada (Rawlusyk, 2018) suggest that high-stakes examinations continue to be used as a dominant mode of assessment, despite growing awareness of the need for more formative and diversified assessment practices.

In this article, we consider whether the enduring popularity of high-stakes summative examinations is justified by empirical evidence through a scoping review of literature on the topic. To our knowledge, this study represents the first comprehensive synthesis of the purported benefits and drawbacks of high-stakes examinations. It offers a summary of the key arguments and an analysis of the evidence that will be of value to university teachers, learning designers, institutional policy makers, and anyone with an interest in assessment in higher education.

While examinations can take many forms, our interest here is in individual, closed-book assessments “undertaken in strict formal and invigilated time-constrained conditions” (Bridges et al., 2002, p. 36), either in-person or via proctored online testing, which occur at the end of a subject (and therefore have a purely summative function). We employ this restricted definition in part because of the prevalence of such examinations in university curricula and in part in acknowledgement that the limitations of examinations identified in the literature and summarized here can be overcome to some extent with more creative design. For example, well-designed open-book and take-home examinations, groupwork examinations, and shorter nested examinations scheduled throughout the semester, potentially all offer an advance on the traditional high-weighted closed-book examination. We consider high-stakes examinations to be those which are strongly consequential for student progression, either due to heavy weighting (often 50% or more of the overall assessment weight associated with a subject), and/or the assignment of ‘hurdle’ status (students must obtain a passing grade in the examination to pass the subject). While high weightings assigned to any assessment task reduce opportunities for students to demonstrate their abilities through more diversified forms of assessment, we argue that the time-constrained, information-restricted and typically written format of the examination make it especially problematic.

In the current higher education landscape, questions about the reliability of assessment methods and their relative vulnerabilities to cheating are highly topical, especially in the wake of the recent availability of generative AI technologies such as ChatGPT. These developments seem likely to spur a call for even greater curricular emphasis on high-stakes examinations. Equally, ongoing concerns about rising costs and growing student numbers mean that universities are under considerable pressure to prioritise assessment methods that are cost effective and efficient, which can result in a default to examinations without sufficient consideration of their pedagogical merits. It is therefore timely to revisit and review the empirical evidence and arguments for and against the use of examinations.

Methods

We conducted a scoping review of the education literature for arguments and evidence for and against the use of high-stakes final examinations in higher education. The purpose of a scoping review is to summarize the body of literature on a given topic and to assess the quality of the evidence (Gómez & Suárez, 2021; Munn et al., 2018). This methodological approach was therefore well suited to the objective of our study. Our selection of studies followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis for Scoping Review (PRISMA-ScR) guidelines which involves four stages: identification of records, screening of records, assessment of eligibility, and the inclusion of studies (Gómez & Suárez, 2021; Liberati et al., 2009). In the first stage, we conducted a search in three databases (Web of Science, Scopus, and Proquest) for all articles written in English and published before July 2023 that contained the term 'high-stakes test,' 'high-stakes exam,' or 'high-stakes assessment' in the title, abstract or keywords, coupled with 'university' or synonyms ('tertiary,' 'higher education,' and 'college') and 'review,' 'evidence,' 'empirical,' or 'analysis.' Our database search returned 406 articles. As part of stage one, in line with the PRISMA guidelines, we also identified articles through other sources, including papers that were known to the authors, articles from reference lists, additional references suggested by reviewers of an earlier version of this paper, and papers that addressed the related topic of 'exam culture.' Through this process, we identified a further 82 articles.

After removing duplicates, in stage two of the process, two of the authors (SF and RM) independently screened the abstracts of the retrieved articles to determine whether they were relevant. Papers were rejected if they were deemed a) off-topic or out of scope; b) not relevant to a higher education teaching and learning setting; c) narrowly focussed on a specific assessment type, assessment instrument or educational context; or d) only peripherally related to the topic. We aimed to include as many claims and sources of evidence for and against examinations as possible and thus empirical grounding was not used as a criterion for inclusion. After independent screening, we reviewed cases where there was a mismatch between the authors in initial assignment of relevance ($n = 8$, 2%) to achieve consensus about inclusion or exclusion (6 papers were excluded and two were retained by consensus). After screening, the 406 articles returned by our database search were narrowed to 40 while the records identified through other sources remained at 82, giving us a total of 122 relevant papers. In the third stage the authors read the full text of the studies selected to confirm that they met the inclusion criteria.

Article coding and analysis

Articles were coded according to emergent themes. Our method of data coding combined an inductive and deductive approach (Braun & Clarke, 2012). Firstly, using an inductive approach, we derived thematic codes from the articles which allowed us to create a long list of codes and to tag each article in our database with the relevant codes. Secondly, we identified the themes, considering areas of commonality between the codes and the key ideas around which the codes clustered (Braun & Clarke, 2012). We refined the themes and combined themes that were conceptually aligned, resulting in seven key themes. Some themes, such as optimal exam design or the prevalence of examination cultures in certain countries, were not helpful in addressing our central question and were therefore omitted from the thematic analysis but are included in the introductory section of this article. We then returned to the papers and applied deductive reasoning to determine whether each paper presented arguments in relation to one or more of the themes as well as whether it offered evidence to support its claims. Each article was classified according to the themes and inserted into a table (Table 1); articles that cross-over more than one theme are listed multiple times. Once the texts were grouped thematically, we summarized and synthesized the arguments and evidence presented within each of the key themes, identified contradictions and inconclusive findings, and highlighted any gaps in the literature.

Results and discussion

Below, we explore each of the above key themes and synthesize the scholarly literature on the topic. Within each theme, we closely examine the empirical evidence relating to both the benefits and drawbacks of high-stakes summative examinations and analyze the key arguments.

Theme 1: memory recall and knowledge retention

Many studies provide empirical evidence that addressing questions in tests and exams can improve memory recall and the retention of information (Butler & Roediger, 2007; Deng et al., 2015; Karpicke & Roediger, 2008; McDaniel et al., 2007; Roediger & Karpicke, 2006). These studies, in the field cognitive psychology, define this phenomenon as ‘the testing effect’ or ‘test-enhanced learning,’ and show that testing produces greater retention than studying. This evidence suggests that within certain disciplines, testing students is of value and is especially useful in courses in which students need to learn a large amount of factual information.

However, the findings of these studies show that the spacing, timing and frequency of effective testing is not aligned with the conditions commonly associated with high-stakes final examinations. Rather, research on the benefits of test-enhanced learning consistently indicates that regular short-answer tests or quizzes taken shortly after the content is taught are of greater value for knowledge retention than single, high-stakes summative examinations (Butler & Roediger, 2007; McDaniel et al., 2007; Santovena-Casal, 2019). Pointing to the power of successive relearning for knowledge retention, Rawson et al (2013) show that memory retrieval is most effective as the number of successful retrievals increase, illustrating the limitations of once-off learning that is encouraged by summative examinations. Low-stakes tests during the semester are therefore a better alternative to high-stakes examinations at the end of the semester in terms of retention.

Despite the benefits of test-enhanced learning, it is also well known that retention of knowledge demonstrated in exams can be short-lived (Greene, 1931; Jones et al., 2015; Rawson et al., 2013). This is likely because the kind of cognitive activity that is required for long-term retention is at odds with that required for rote-learning facts, which tends to be the dominant form of activity required exam preparation. Content analyses of examinations in university science and medicine courses, for example, have shown that most questions merely test the isolated recall of factual knowledge (Ramsden, 1992). Without a deeper understanding of the relevance, context and application of concepts, abstract ideas are easily forgotten. Examinations may also undermine long-term knowledge acquisition because they emphasise extrinsic reward (Kuhbandner et al., 2016), enticing students to memorise facts so that they can perform well on an examination rather than engage in deeper learning.

Table 1 Listing of broad themes and the publications corresponding to each theme

Broad theme	Publications
1: Memory recall and knowledge retention	Greene, 1931 Roediger & Karpicke, 2006 Butler & Roediger, 2007 McDaniel et al., 2007 Karpicke & Roediger, 2008 Roediger & Butler, 2011 Rawson et al., 2013 McConnell et al., 2015 Deng et al., 2015 Kuhbandner et al., 2016 Santovena-Casal, 2019
2: Student motivation and learning	Smith, 1991 Ramsden, 1992 Harlen & Deakin Crick, 2003 Jones et al., 2003 Marchant & Paulson, 2005 Wise & DeMars, 2005 Gijbels & Dochy, 2006 Trotter, 2006 Wise, 2009 Berliner, 2011 Hartwig & Dunlosky, 2012 DeWitt et al., 2013 Surgenor, 2013 Wang & Brown, 2014 Williams, 2014 Zhan & Andrews, 2014 Harland et al., 2015 Wass et al., 2015 Kuhbandner et al., 2016 Benediktsson & Ragnarsdóttir, 2020 Biggs et al., 2022 Wang et al., 2022
3. Authenticity and real-world relevance	Gibbs & Lucas, 1997 Knight, 2002 Knight & Yorke, 2003 Boud & Falchikov, 2006 Williams, 2008 Van Bergen & Lane, 2014 Williams, 2014 Durning et al., 2016 Stopar & Ilc, 2017 Boud, 2018 Villaroel et al., 2020 Choi & Chun, 2022

Table 1 (continued)

Broad theme	Publications
4. Validity and reliability	Hartog & Rhodes, 1936 Cronbach, 1971 Cox, 1973 Frederiksen & Collins, 1989 Entwistle & Entwistle, 1991 Messick, 1992 Sambell et al., 1997 Sternberg, 1997 Smith & Fey, 2000 Knight, 2002 Haertel, 2006 Mason, 2007 Stobart, 2009 Shaw et al., 2012 Caines et al., 2014 Kellaghan & Greaney, 2019 Eweda et al., 2020
5. Academic misconduct and contract cheating	Sheard & Dick, 2003 McCabe, 2005 Sutherland-Smith, 2008 Van Bergen & Lane, 2014 Potaka & Huang, 2015 Baird & Clare, 2017 Lancaster & Clarke, 2017 Bretag, et al., 2019a Bretag, et al., 2019b Dawson, 2020 Ellis et al., 2020 Ali & Alhassan, 2021 Awdry, 2021 Hill et al., 2021 Peh et al., 2021 Raman et al., 2021 Reedy et al., 2021 Crossley, 2022

Table 1 (continued)

Broad theme	Publications
6. Stress, anxiety and wellbeing	Hembree, 1988 Wolf & Smith, 1995 Maes et al., 1998 Weekes et al., 2006 Zhang et al., 2011 Fernández-Castillo & Caurcel, 2015 Sommer & Arendasy, 2015 Lotz & Sparfeldt, 2017 Fernández-Castillo & Caurcel, 2019 Vogel & Schwabe, 2016 Franke, 2018 Hamzah et al., 2018 Von Der Embse et al., 2018 Shean, 2019 Fejes et al., 2020 ShayesteFar, 2020 Monrad et al., 2021 Roos et al., 2021 Fawaz & Lee, 2022 Högberg & Horn, 2022 Theobald et al., 2022 Fang et al., 2023
7. Fairness and Equity	Marchant & Paulson, 2005 Woodfield et al., 2005 Rask & Tiefenthaler, 2008 Klenowski, 2009 Claypool & Preston, 2013 Karami, 2013 Richardson, 2015b Uy et al., 2015 Gliatto et al., 2016 De Paola & Gioia, 2016 Klenowski, 2016 Ballen et al., 2017 Niessen et al., 2019 Salehi et al., 2019 Trumbull & Nelson-Barber, 2019 Bordbar, 2020 Jackson et al., 2020 Nieminen & Tuohilampi, 2020 Burgoyne et al., 2021 Mehrazmay et al., 2021 Preston & Claypool, 2021 Meeks et al., 2022 Nieminen, 2022 Tai et al., 2022 Crossley, 2022

Thus, while there is evidence that testing may assist with knowledge retention, high-stakes summative examinations are ultimately ill-suited to deliver benefits of test-enhanced learning because they do not involve repetition, occur towards the end of the learning process, involve large amounts of content that is difficult to retain, and encourage rote learning.

Theme 2: student motivation and learning

A potential pedagogical benefit of examinations lies in the role they can play in motivating students to study. Studies examining self-reported levels of motivation have found (perhaps unsurprisingly) that student motivation to study is higher for high-stakes assessments than low-stakes assessments, and that motivation is a positive predictor of outcomes (Wise, 2009; Wise & Demars, 2005). Study habits are also known to be an important factor in retrieval, retention, and student achievement, including the practices of self-testing, rereading, and scheduling of study (Hartwig & Dunlosky, 2012; Roediger & Butler, 2011). However, if student motivation to study stems from a desire to rote-learn information to perform well on an examination, extrinsic motivation is activated, but not intrinsic or autonomous motivation, which has been shown to be far more important for student learning (Ryan & Deci, 2000) and long-term memory acquisition (Kuhbandner et al., 2016). In their systematic review of research on the impact of testing on students' motivation for learning, Harlen and Deakin Crick (2003) found that high-stakes examinations reduce intrinsic motivation and have a negative impact on student self-regulation. Similarly, studies reporting on student preferences find that students do not perceive high-stakes examinations to be motivating or beneficial for their learning (Benediktsson & Ragnarsdóttir, 2020; Gijbels & Dochy, 2006; Wang & Brown, 2014). Students prefer a broad range of assessment tasks spread throughout the semester (Surgenor, 2013; Trotter, 2006; Wass et al., 2015), provided that the use of continuous assessment is not so frequent that students are over-assessed (Harland et al., 2015; Wass et al., 2015). There are also limitations to the efficacy of exam study as a learning method since students tend to adopt traditional methods of studying such as memorization-related activities (Biggs et al., 2022) or reviewing past exams, rather than application-related activities or more authentic modes of study. For example, Zhan and Andrews (2014) found that English language students in China developed their listening skills for a high-stakes examination by listening to past examination recordings, rather than by listening to authentic English media as was intended by the designers of the exam.

A significant pedagogical disadvantage of summative high-stakes examinations identified in the literature is their encouragement of superficial or 'strategic' learning whereby students focus only on studying content that has the potential to enable them to gain higher grades (Williams, 2014). The term *backwash,* coined by Elton (1987), refers to the effect that assessment has on student learning, which can take either negative or positive forms. As Biggs et al., (2022, p. 188) observe, 'negative backwash always occurs in an exam-dominated system' in which 'strategy becomes more important than substance.' Examinations therefore tend to encourage surface-level learning (DeWitt et al., 2013; Gibbs, 1992), as discussed in Theme 1.

In part, such limitations are an outcome of poorly designed examinations that purely test the recall of information. Examinations can be designed to promote higher levels of thinking, such as by posing questions that demand in-depth analysis and synthesis, and by placing questions within contexts relevant to the field being studied (McConnell et al., 2015; Villarroel et al., 2019). The format of the examination also determines the range of skills and knowledge that can be fostered and assessed. For example, open book and take-home examinations that allow students to engage with materials and integrate knowledge have a greater potential to measure the application and synthesis of knowledge than closed book invigilated examinations (Durning et al., 2016), and to assess authentic tasks using real-world scenarios (Deneen, 2020). However, even when they are designed to promote higher order thinking, there are limitations as to the kinds of skills examinations can assess, given that their format is almost exclusively written.

A further impediment to student learning results from the responses of teachers to high-stakes examinations, which encourage educators to 'teach to the test' (Marchant & Paulson, 2005; Smith, 1991) and spend large amounts of time on test preparation (Jones et al., 2003). Berliner (2011) writes of the way in which examinations lead to curriculum narrowing, which he describes as a rational and inevitable response to high-stakes testing. Teachers will naturally align their curriculum with assessment, narrowing the content they teach to what will (and can) be assessed within the final examination. This practice is especially problematic when curriculum narrowing favors activities that call for low-level cognitive processes. Surface learning is a consequence of high-stakes testing since examinations are limited in their capacity to measure higher-order thinking, and teachers' structure their curriculum accordingly (Berliner, 2011; Jones et al., 2003; Smith, 1991).

Therefore, while high-stake examinations undoubtedly motivate students to invest time and prepare for the assessment, the nature of this investment is often geared towards maximising grades rather than learning, and at odds with demonstration of high-order thinking and skills.

Theme 3: authenticity and real-world relevance

Some argue that examinations reflect real-life situations in the workplace, especially in fields such as the medical and health professions, where information and facts must often be recalled, and decisions made, under time-pressure and without recourse to materials (Durning et al., 2016; Van Bergen & Lane, 2014). This is considered an essential skill by employers in a limited set of academic disciplines. However, it is debatable whether this rationale applies to most other disciplines, and whether performance in an examination is actually a useful proxy for performance under pressure in the workplace, since the artificially constructed nature of the exam format is unlikely to authentically reflect a genuine workplace situation.

Indeed, a recurring argument against the use of examinations in the literature is their lack of authenticity and limited capacity to foster the kinds of skills and knowledge students will need in their future careers, which are more likely to require skills such as critical thinking, problem solving and communication, and the application of knowledge than the ability to recall facts (Boud & Falchikov, 2006; Gibbs & Lucas, 1997; Williams, 2008, 2014). While improving the authenticity of examinations by designing questions that reflect real-life situations and require evaluative judgement has been shown to support deeper approaches to learning (Villarroel et al., 2020), even when designed effectively, examinations are limited in their capacity to support the principles of authentic assessment. Examinations have especially limited capacity to foster and assess listening and communication skills (Choi & Chun, 2022; Stopar & Ilc, 2017), two of the most valued skills employers report wanting in university graduates (Bauer-Wolf, 2019). Assessment tasks that foster these skills, such as presentations, debates, peer-peer learning activities, and inquiry-based group projects all offer potentially valuable opportunities for authentic learning that are reduced when there is a heavy emphasis on high-stakes examinations.

It is also highly unlikely that summative exams will prepare students to be life-long learners, because they position students as passive recipients of feedback without encouraging them to judge the quality of their work, or to apply the feedback they receive (Boud, 2018; Boud & Falchikov, 2006). The results of summative examinations are final and preclude any opportunity for students to learn from mistakes or improve their performance (Knight, 2002). The capacity to reflect upon, critically appraise and improve one's own work are likely to be essential to students' future lives and careers. Examination formats are therefore poorly aligned with the imperative for students to develop the capacity to self-assess and to receive and implement feedback. The provision of quality and timely feedback is an essential element of formative assessment, as is the active role of the student in engaging with and implementing the feedback they receive (Henderson et al., 2020; Winstone & Carless, 2020), to ensure continuous improvement and enhanced performance. The capacity for formative assessments to better allow students to demonstrate their ability is evidenced by the numerous studies showing that student performance is improved in assessments that take place outside of the examination context (Bridges et al. 2002; Richardson, 2015a; Simonite, 2003).

The high-stakes nature of most summative examinations also discourages students from adopting an experimental approach to learning, which is a key desired graduate capability. Making mistakes and experiencing misconceptions are an essential part of learning (Metcalfe, 2017; Verdake et al., 2017), which is why it is important that sufficient opportunities are provided for students to engage in low-stakes (or no-stakes) assessments early in a subject, and given sufficient opportunity to use 'error detection' (Biggs et al., 2022, p. 186) as the basis for correcting and learning from their mistakes.

In summary, the information-limited, high-pressure context of high-stakes examinations is far removed from most authentic workplace situations. The restricted format of such assessments and their summative nature furthermore limits opportunities to assess key generic skills and discourages an experimental approach to learning through iterative cycles of feedback and improvement.

Theme 4: validity and reliability

When assessments are highly consequential for selection, progression, recognition or certification, it is self-evident that students, teachers and other stakeholders must have confidence that the instrument has high validity and reliability. A simple definition of the validity of an assessment is that it assesses what it claims to assess (and thus allows meaningful, accurate, and appropriate inferences to be made from its scores; Messick, 1992)). However, validity is a complex and multifaceted construct with dimensions ranging from *construct validity* (whether the test measures the concept it was intended to measure), to *content validity* (whether the test includes a representative set of all aspects of the construct), *face validity* (whether the content of a test seems appropriate to its aims), *criterion validity* (whether test scores correlate with functional behaviours the test seeks to measure), and *consequential validity* (whether the test has potential or actual positive or negative consequences for teaching and learning; Messick, 1992).

High-stakes university examinations are problematic from the perspective of validity for several reasons. Firstly, while there is well-developed literature on validity testing, validation of high-stakes examinations administered in university courses is neither required nor routinely undertaken (indeed, even formal validation of large national high-stakes tests Mason, 2007; Stobart, 2009) is uncommon (Messick, 1992)). University examinations are typically set by academic staff who have extensive disciplinary knowledge but scant expertise in maximizing or evaluating the validity of their examinations (an exception is the clinical sciences, where 'blueprinting' is often used to improve the content validity of examinations; Eweda et al., 2020). University assessment policies often require staff to prepare 'supplementary' examination papers for students who are unable to sit the initial examination, but formal comparison of the validity equivalence of exam variants is also seldom undertaken. The rarity of formal validation is likely due to the complexity of the concept of validity, the lack of agreed approaches to validation, the multiple methods needed to provide comprehensive validity evidence, and the effort required to undertake such endeavours (Shaw et al., 2012).

Secondly, because validity is not an inherent characteristic of a test, but of the test in the context in which it is used (Cronbach, 1971), even if a particular test is found to be valid for one purpose (e.g., norm-referenced comparison of accomplishment across students in a cohort), the same test may not be valid for another purpose (e.g. as a pass/fail decision-making tool; Smith & Fey, 2000). Sternberg (1997) found that summative assessments which had high predictive validity in relation to achievements at points in an undergraduate degree were often moderate or poor predictors of subsequent career achievement.

Thirdly, depending on their perspectives, teachers and students may have very different perceptions about whether a test is valid and fair (Caines et al., 2014). As discussed in Theme 2, high-stakes examinations have been extensively criticised for their poor consequential validity. Frederiksen and Collins (1989) defined *systemically valid* tests as those that drive curricular and instructional changes in education systems that foster the development of the cognitive traits that the tests are designed to measure. They suggested that high-stakes examinations led to an undesirable narrowing of what is taught due to excessive focus on meeting test requirements. Sambell et al (1997) reported that students had very negative views of traditional 'unseen' (closed-book) examinations, perceiving that the strategies required to perform well on such examinations encouraged shallow or poor learning, distorting the quality of their learning. Entwistle and Entwistle (1991) also found that examinations both distorted students' efforts to achieve genuine understanding and that examination questions often did not tap conceptual understanding.

Examinations have also long been criticized for their low reliability (Hartog & Rhodes, 1936), defined as their ability to produce consistent and reproducible results. Low reliability of examination performance can result from a range of factors, related to examinees, examiners, the subject being examined, the test items, and how the examination is scored (Haertel, 2006). For instance, the same examinee may perform differently on an examination because of their psychological or physical health, the conditions of the examination space, or their familiarity with the test items selected for the exam. There are also many factors that can affect the performance and judgement of the examiner, including bias, inconsistency, or rater drift (Cox, 1973; Hartog & Rhodes, 1936; Kellaghan & Greaney, 2019; Knight, 2002). Finally, the number of test items, whether rubrics are used, and how many items are included in the examination can all affect how it is marked. Collectively, these factors significantly reduce confidence in the reliability of examination scores.

Issues of validity and reliability are arguably not unique to examination assessments. However, the absence of a culture of validation, the difficulty of achieving high validity, the evidence for low consequential validity, and the myriad factors that can affect the reliability of examination performance means that there is a troubling lack of validity and reliability evidence underpinning the culture of high-stakes, 'one-chance' examinations.

Theme 5: academic misconduct and contract cheating

One of the most common arguments in favour of high-stakes summative examinations is the belief that they are more effective than other forms of assessment at preventing contract cheating. This is one reason why invigilated closed-book examination formats are favoured: students complete these assessments in a tightly controlled environment, providing photo ID and completing the examination in an open public space under close observation. This ought to minimize opportunities for cheating and plagiarism (Crossley, 2022; Van Bergen & Lane, 2014).

However, it is evident that even these tightly controlled contexts do not provide reliable protection against academic misconduct and cheating (Lancaster & Clarke, 2017). Indeed, contract cheating in relation to examinations appears to be prevalent, involving behaviors that range from collusion to impersonation (Bretag et al., 2019a), facilitated by the apparent ease with which university student identification cards can be forged (Potaka & Huang, 2015). Sheard and Dick (2003) estimated that the frequency of cheating in examinations in a cohort of graduate students in IT courses approached ten percent, while in McCabe's (2005) study of 64,000 North American university students, over one-third admitted to some form of exam cheating. Bretag et al (2019a) found in a large survey of Australian universities that students participated in undetected cheating in invigilated examinations at higher rates than any other type of cheating, including contract cheating in written assessments. The frequency of academic misconduct in examinations undoubtedly increased with the move to online learning during the COVID-19 pandemic (Hill et al., 2021; Peh et al., 2021; Reedy et al., 2021), and challenges around online proctoring (Raman et al., 2021), mean that there is only minimal capacity to ensure academic integrity in closed-book online examinations. Noting that students perceive contract cheating to be most likely for heavily weighted assignments, Bretag et al., (2019b, pp. 685–686) suggest that "examinations provide universities and accrediting bodies with a false sense of security" and that "an over reliance on examinations, without a thorough and comprehensive approach to integrity, is likely to lead to more cheating, not less."

Effective exam design and delivery can minimize opportunities for contract cheating; for example, examination papers should not be reused, online tests should not be unsupervised, and low-level or 'one right answer' tasks and questions should be avoided (Dawson, 2020). However, concerns about academic misconduct can more effectively be alleviated through alternative forms of assessment that limit or remove the potential for cheating. The increasing prevalence of "assignment outsourcing" by ghost writers and essay mills is well known (Ali & Alhassan, 2021; Awdry, 2021), and the essay format is likely to continue to cause concerns relating to academic misconduct, especially considering the increased uses of AI software such as ChatGPT. However, careful assessment design may help combat or reduce opportunities and incentives for cheating in a range of ways (Baird & Clare, 2017). For instance, assessment tasks or questions that ask students to reflect or draw on personal circumstances or experiences (Sutherland-Smith, 2008), local contexts or environments, or assessment tasks that are conducted within a specific class or tutorial activity should generally be more difficult to procure from external sources than standard essays on common topics. Similarly, where tasks involve repeated contributions (reflective journals and blogs), audit trails of progress, or other forms of 'authentic' assessment, they ought to be difficult or costly to obtain from external providers. Finally, authorship of some assessment tasks such as vivas, individual or group oral presentations, or video presentations can be verified with a relatively high degree of confidence. It is nevertheless important to recognise that ultimately, with the possible exception of labor-intensive formats like interviews or vivas, very few assessment tasks are immune to outsourcing (Bretag, et al., 2019b; Ellis et al., 2020).

In summary, several large empirical studies challenge the widespread belief that invigilated high-stakes examinations offer better security against contract cheating (and instead suggest that they may be particularly vulnerable). While effective exam design and delivery measures can reduce cheating opportunities, academic integrity concerns alone do not provide compelling grounds for maintaining an overreliance on high-stakes examinations. Educational institutions should explore a broader range of assessment methods that better align with the evolving challenges of academic misconduct in the digital age.

Theme 6: stress, anxiety, and wellbeing

High-stakes examinations have long been associated with psychological distress and anxiety (Hembree, 1988; Kellaghan & Greaney, 2019; Lotz & Sparfeldt, 2017). This issue has come to the fore in recent years with an increased focus on the role that curriculum and assessment design play in supporting student mental wellbeing (Baik et al., 2019; Slavin et al., 2014). Physiological measures of stress including cardiovascular parameters and stress hormones have been shown to be higher during examination periods compared to outside these periods (Fejes et al., 2020; Maes et al., 1998; Weekes et al., 2006; Wolf & Smith, 1995; Zhang et al., 2011). Students also self-report higher levels of anxiety during examination periods (Ballen et al., 2017; Högberg & Horn, 2022; Zhang et al., 2011); a reliable correlate of physiological stress (Roos et al., 2021). Studies have found a correlation between the competition amongst peers promoted by high-stakes exams and negative mental health impacts, including emotions of shame, self-loathing (Fang et al., 2023), and suicidal ideation, especially in female students (Fawaz & Lee, 2022).

There is some dispute regarding whether distress and anxiety negatively impact on examination performance. Many studies have found that anxiety does have a negative impact on performance (Hembree, 1988; Stenlund et al., 2018; Von Der Embse et al., 2018; Wolf & Smith, 1995), while others have found that anxiety does not have a statistically significant effect on performance (Monrad et al., 2021; Sommer & Arendasy, 2015). Some even argue that examination anxiety is useful as it promotes study and preparation (Hamzah et al., 2018) and can increase performance (Shean, 2019). There is also disagreement on whether stress interferes with the retrieval of previously learned knowledge, with some studies finding that stress impairs memory retrieval (Vogel & Schwabe, 2016), while others find it does not (Theobald et al., 2022).

Such contradictory findings hinder a conclusive understanding of the relationship between anxiety and performance. Nevertheless, there is evidence to suggest that a range of other factors associated with anxiety may have an impact on the capacity for students to perform well. For example, anxiety has been found to correlate negatively with motivation, which has a direct effect on achievement (ShayesteFar, 2020). Students more prone to examination anxiety are also more likely to have lower self-esteem and sleep less during examination periods, which reduces concentration (Fernández-Castillo & Caurcel, 2019). It is likely that higher-weighted examinations have stronger negative impacts on students' wellbeing, due to the increased perception of consequences of the outcomes (Franke, 2018; Salehi et al., 2019; Wolf & Smith, 1995). There are also a range of cultural and genetic factors that exacerbate experiences of examination stress (Zhang et al., 2011), meaning that students will be affected unequally, as we discuss in more detail below. Finally, there is evidence that assessments which are perceived by students as threatening and which provoke anxiety may drive students to adopt surface rather than deep approaches to learning (Gibbs, 1992; Ramsden, 1992).

As outlined above, there is substantial evidence that examinations cause elevated distress and anxiety. Although the impact of examination anxiety on student performance is inconclusive, the proven adverse effects of examinations on student mental health and wellbeing is concerning, as is the negative impact of examination anxiety on student motivation.

Theme 7: fairness and equity

There is some anecdotal opinion that examinations are a fair form of assessment, a 'test of truth' that allows students to compete and demonstrate their individual capabilities on equal footing (Crossley, 2022). However, such opinions are not supported by the empirical evidence. It is known that students perform differently under time pressure (De Paola & Gioia, 2016), and there is considerable evidence that examinations have the potential to generate academic inequity due differential performances based on gender (Ballen et al., 2017; Mehrazmay et al., 2021; Rask & Tiefenthaler, 2008; Salehi et al., 2019), socio-economic status (Gliatto et al., 2016; Jackson et al., 2020; Uy et al., 2015), race and ethnicity (Claypool & Preston, 2013; Klenowski, 2009, 2016; Marchant & Paulson, 2005; Preston & Claypool, 2021; Richardson, 2015a; Trumbull & Nelson-Barber, 2019), and disability (Meeks et al., 2022; Nieminen, 2022; Nieminen & Tuohilampi, 2020; Tai et al., 2022). This substantial body of literature on the equity implications of high-stakes examinations provides compelling evidence that examinations can disadvantage marginalized groups and contribute to academic inequity, which intersects with impacts on wellbeing and student learning.

There are a range of studies within the STEM disciplines that suggest examinations differentially affect students based on their gender, finding that women tend to suffer from higher levels of assessment anxiety leading to lower wellbeing and reduced concentration during an examination, and resulting in lower performance (Fernández-Castillo & Caurcel, 2015, 2019; Roos et al., 2021; Salehi et al., 2019), an effect that may be stronger at introductory levels of university (Ballen et al., 2017; Salehi et al., 2019). A study by Ballen et al. (2017) found that women in an introductory biology course underperformed on examinations compared to their male counterparts but outperformed them on combined non-examination methods of assessment. Salehi et al. (2019) argue that the use of high-stakes examinations as a primary assessment method in the STEM disciplines, especially in introductory level courses, imposes a "gender penalty" on female students that may prevent them from advancing in the discipline. Gendered differences have also been found in how students respond to outcomes, with studies suggesting that women may be more likely than men to leave a course after receiving low marks on an introductory course (Rask & Tiefenthaler, 2008; Salehi et al., 2019).

However, there is some debate about the relationship of gender to performance and preferences across modes of assessment, and not all studies provide conclusive evidence of gender bias. For example, a study at the University of Sussex (Woodfield et al., 2005) found that women outperformed men by a small margin on both coursework assignments and final examinations and that students of both sexes performed better on coursework than examinations. Some studies have found no evidence of differential performance on the basis of gender (Karami, 2013; Niessen et al., 2019), while others have found that different components of examinations and styles of questions favour one gender over the other (Bordbar, 2020; Burgoyne et al., 2021), suggesting that examination designs might have more significant equity implications than the assessment method itself. There are a range of intersecting factors and independent variables that render a definitive conclusion about whether exams perpetuate gender bias problematic. However, there is sufficient evidence in the literature to suggest that there may be correlations between assessment modes and gendered styles of learning that warrant consideration when designing assessment tasks.

In addition to potential gender biases, examinations may be biased towards Western students, with equity implications for international students from non-Western countries and for Indigenous students. Richardson (2015a, b) suggests it is likely that the under-attainment of students from ethnic minorities is connected to assessment methods, while many have argued that examinations disadvantage Indigenous students (Claypool & Preston, 2013; Klenowski, 2009; Preston & Claypool, 2021; Trumbull & Nelson-Barber, 2019), as they tend to promote Western intellectual knowledge and values by supporting the view that knowledge can be given, accumulated, and tested in a linear manner. Recent scholarship on inclusive assessment design further argues that examinations fail to meet the needs of student diversity, especially with respect to students with disabilities (Nieminen, 2022; Nieminen & Tuohilampi, 2020; Tai et al., 2022). If examinations favour students from certain groups (male, western, able-bodied) as this research suggests, their reliability and validity as measures of student achievement also comes into question.

Overwhelmingly, the research suggests that 'once-chance,' time-pressured final examinations have exclusionary effects and disadvantage marginalised student groups. Alternative forms of assessment that allow for more diverse formats (including non-written formats), as well as formative assessments that offer more support for students, are therefore better aligned with the principles of inclusive assessment design.

Conclusion

Our scoping review of the literature suggests that the current heavy reliance on high-stakes final examinations in many university subjects is poorly justified by the balance of empirical evidence, and that traditional examinations (closed-book, individual, invigilated, time-constrained, summative, final, and high-stakes) have limited pedagogical value. However, the evidence on the benefits of test enhanced learning for memory recall and knowledge retention (Theme 1) along with the role that examinations can play in motivating students to study (Theme 2), indicate that well-designed examinations in a revised format do have a role to play in the curriculum in some subjects, especially when they are formative and low-weighted. To be beneficial to student learning, the format of the examination must engage students in high-order skills, which can potentially be achieved in open-book and take-home examinations, short examinations, or tests scheduled throughout the semester that can build student learning over time, and in groupwork examinations, which can be employed to engage students in collaborative learning tasks. Regardless of their format, it is imperative that examinations are well designed for both pedagogical and security reasons. For example, short-answer questions as well as context-rich multiple-choice questions that require the application of knowledge can enhance learning relative to multiple-choice questions that require the recall of facts (McConnell et al., 2015). To reduce opportunities for contract cheating, low-level or 'one right answer' tasks and questions should be avoided (Dawson, 2020).

While the evidence presented in this paper within Themes 1 and 2 suggests that well-designed examinations can be of value under the kinds of conditions outlined above, the pedagogical drawbacks of examinations across all other themes illustrates that it is highly problematic when high-stakes final examinations dominate the curriculum. The artificially constructed nature of the examination format limits their authenticity and real-world relevance (Theme 3) and prevents opportunities for students to self-asses, and implement feedback, which are fundamental to becoming life-long learners. The absence of formal validation, the difficulty of achieving validity, and the low-reliability of examination performance (Theme 4) raises serious concerns about the role examinations frequently play as highly consequential measurements of student performance and capacity. Although guarding against academic misconduct and contract cheating (Theme 5) are commonly-cited reason for the ongoing use of examinations, the evidence shows that contract cheating in examinations is not only prevalent but may be even more pronounced than in other forms of assessment. Increasing the use of invigilated final examinations will not fix this problem. Instead, universities need a comprehensive approach to integrity that includes careful assessment design and forms of authentic assessment that mitigate against the potential for cheating.

The role that high-stakes examinations play in contributing to increased stress and anxiety and decreased student wellbeing (Theme 6) is one of the most troubling findings of this review. The literature provides considerable evidence to show that examinations have adverse effects on student physical and mental health and demonstrates the negative impacts of examination anxiety on student motivation, concentration, and deep approaches to learning. Even if, as some studies claim, anxiety can lead to increased study and performance (and many studies dispute this claim), such potential gains need to be weighed carefully against the negative impacts on wellbeing. Moreover, students have differing capacities to perform effectively under stress and time-pressure, and differential performance is also likely on the basis of gender, socio-economic status, race, ethnicity, and disability. The potential for examinations to disadvantage marginalized student groups and perpetuate educational and social inequity (Theme 7) is especially concerning when they are high-weighted and have significant consequences for students' lives and careers.

The pronounced lack of empirical evidence for the pedagogical benefits of high-stakes examinations suggests that they are employed primarily for reasons related to cost, efficiency, practicality, scalability, and administrative convenience. However, we question whether these reasons remain valid in the contemporary higher education landscape where many advances in assessment design are already well established, and others are emerging. There are promising examples of educational technology that can assist with the administrative burden of distributing and grading assessments other than high-stakes examinations at scale. Examples include platforms that support peer assessment (Søndergaard & Mulder 2012), social annotation (Miller et al., 2018), personalizing feedback through the use of digital recordings (Ryan et al., 2021), and automated feedback and grading (Cavalcanti et al., 2021; Hegarty-Kelly & Mooney, 2021; Kumar & Boulanger, 2021). Programatic assessment also offers an approach to assessment design that has the potential to both increase assessment security (Dawson, 2020) and to reduce the reliance on high-stakes summative examinations by diversifying assessment methods across the curriculum of an entire program (Baartman et al., 2022; Heeneman et al., 2021). While not trivial to implement, a programmatic approach allows for intentional emphasis on low-weighted assessments in the foundational years of a degree, placing more emphasis on assessment tasks that foster the development of student capabilities and cohort connections. Examinations could then be employed at key moments for accreditation purposes (although other methods might also fulfil this role), but final summative examinations would no longer be the default assessment mode.

The use of high-stakes examinations becomes particularly problematic when they dominate the curriculum at the expense of other valuable forms of assessment. It is essential that students are provided with the opportunity to engage with a broad range of assessment tasks, to develop their learning and build diverse skills that align with desired graduate outcomes and promote a culture of life-long learning. Variation in assessments is critical to allow students to build, apply, and demonstrate different kinds to skills; to foster skills in diverse areas such as self-assessment, inquiry-based learning, communication, and teamwork; and to be rewarded for originality rather than conformity (Ramsden, 1992). Designing assessment practices that encourage continuous and high-quality learning while supporting student wellbeing is a challenging but important task that requires creative and innovative approaches that move beyond an over-reliance on high-stakes summative examinations.

Declarations

Conflict of interest The authors declare that there exists no competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

References

Ali, H. I. H., & Alhassan, A. (2021). Fighting contract cheating and ghostwriting in Higher Education: Moving towards a multidimensional approach. *Cogent Education, 8*(1), 1885837. https://doi.org/10.1080/2331186X.2021.1885837

Awdry, R. (2021). Assignment outsourcing: Moving beyond contract cheating. *Assessment & Evaluation in Higher Education, 46*(2), 220–235. https://doi.org/10.1080/02602938.2020.1765311

Baartman, L., van Schilt-Mol, T., & van der Vleuten, C. (2022). Programmatic assessment design choices in nine programs in higher education. *Frontiers in Education, 7*, 931980. https://doi.org/10.3389/feduc.2022.931980

Baik, C., Larcombe, W., & Brooker, A. (2019). How universities can enhance student mental wellbeing: The student perspective. *Higher Education Research & Development, 38*(4), 674–687. https://doi.org/10.1080/07294360.2019.1576596

Baird, M., & Clare, J. (2017). Removing the opportunity for contract cheating in business capstones: A crime prevention case study. *International Journal for Educational Integrity, 13*(1), 6. https://doi.org/10.1007/s40979-017-0018-1

Ballen, C. J., Salehi, S., & Cotner, S. (2017). Exams disadvantage women in introductory biology. *PLoS ONE*, *12*(10). https://doi.org/10.1371/journal.pone.0186419

Bassey, M. (1971). *The assessments of students by formal assignments*. New Zealand University Students Association. http://www.tandfonline.com/doi/abs/10.1080/02602938.2014.919628. Accessed 11/27/2023

Bauer-Wolf, J. (2019). Survey: Employers wants "soft skills" from graduates. *Inside Higher Ed*. https://www.insidehighered.com/quicktakes/2019/01/17/survey-employers-want-soft-skills-graduates. Accessed 11/27/2023

Benediktsson, A. I., & Ragnarsdóttir, H. (2020). Immigrant students' experiences of assessment methods used in Icelandic universities. *Multicultural Education Review, 12*(2), 98–116. https://doi.org/10.1080/2005615X.2020.1756090

Berliner, D. (2011). Rational responses to high stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education, 41*(3), 287–302. https://doi.org/10.1080/0305764X.2011.607151

Biggs, J. (2001). The reflective institution: Assuring and enhancing the quality of teaching and learning. *Higher Education, 41*(3), 221–238. https://doi.org/10.1023/A:1004181331049

Biggs, J. B., Tang, C. S., & Kennedy, G. (2022). *Teaching for quality learning at university* (5th ed.). Open University Press.

Bordbar, S. (2020). Investigating gender-biased items in a high-stakes language proficiency test: Using the Rasch model measurement. *Applied Linguistics Research Journal*. https://doi.org/10.14744/alrj.2020.73645

Boud, D. (Ed.). (2018). *Developing evaluative judgement in higher education*. Routledge.

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22*(2), 151–167. https://doi.org/10.1080/713695728

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment and Evaluation in Higher Education, 31*(4), 399–413. https://doi.org/10.1080/02602930600679050

Boud, D. (1995). *Enhancing learning through self-assessment* (1st ed.). Routledge. https://doi.org/10.4324/9781315041520

Braun, V., & Clarke, V. (2012). Chapter 4: Thematic analysis. In *APA Handbook of research methods in psychology* (Vol. 2, pp. 37–71). American Psychological Association.

Bretag, T., Harper, R., Burton, M., Ellis, C., Newton, P., Rozenberg, P., Saddiqui, S., & van Haeringen, K. (2019a). Contract cheating: A survey of Australian university students. *Studies in Higher Education, 44*(11), 1837–1856. https://doi.org/10.1080/03075079.2018.1462788

Bretag, T., Harper, R., Burton, M., Ellis, C., Newton, P., van Haeringen, K., Saddiqui, S., & Rozenberg, P. (2019b). Contract cheating and assessment design: Exploring the relationship. *Assessment & Evaluation in Higher Education, 44*(5), 676–691. https://doi.org/10.1080/02602938.2018.1527892

Bridges, P., Cooper, A., Evanson, P., Haines, C., Jenkins, D., Scurry, D., Woolf, H., & Yorke, M. (2002). Coursework marks high, examination marks low: Discuss. *Assessment & Evaluation in Higher Education, 27*(1), 35–48. https://doi.org/10.1080/02602930120105045

Burgoyne, A. P., Mashburn, C. A., & Engle, R. W. (2021). Reducing adverse impact in high-stakes testing. *Intelligence, 87*, 101561. https://doi.org/10.1016/j.intell.2021.101561

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*(4–5), 514–527. https://doi.org/10.1080/09541440701326097

Caines, J., Bridglall, B. L., & Chatterji, M. (2014). Understanding validity and fairness issues in high-stakes individual testing situations. *Quality Assurance in Education, 22*(1), 5–18. https://doi.org/10.1108/QAE-12-2013-0054

Carless, D., Bridges, S. M., Chan, C. K. Y., & Glofcheski, R. (Eds.). (2017). *Scaling up assessment for learning in higher education* (Vol. 5). Springer Singapore. https://doi.org/10.1007/978-981-10-3045-1

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence, 2*, 100027. https://doi.org/10.1016/j.caeai.2021.100027

Chen, Q., Hao, C., & Xiao, Y. (2020). When testing stakes are no longer high: Impact on the Chinese College English learners and their learning. *Language Testing in Asia, 10*(1), 6. https://doi.org/10.1186/s40468-020-00102-5

Choi, Y., & Chun, J. (2022). Test review: French examination of the College Scholastic Ability Test in Korea. *Language Testing in Asia, 12*(1), 49. https://doi.org/10.1186/s40468-022-00199-w

Claypool, T. R., & Preston, J. P. (2013). Redefining learning and assessment practices impacting aboriginal students: Considering aboriginal priorities via aboriginal and Western worldviews. *In Education*, *17*(3). https://doi.org/10.37119/ojs2011.v17i3.74

Cox, R. J. (1973). Traditional examinations in a changing society. *Universities Quarterly, 27*(2), 200–216. https://doi.org/10.1111/j.1468-2273.1973.tb00426.x

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education.

Crossley, M. (2022). Merlin Crossley makes the case for exams. *Camups Morning Mail*. https://campusmorningmail.com.au/news/merlin-crossley-makes-the-case-for-exams/. Accessed 27 Nov 2023.

Dawson, P. (2020). Structuring and designing assessment for security. In *Defending assessment security in a digital world: Preventing E-cheating and supporting academic integrity in higher education* (1st ed.). Routledge. https://doi.org/10.4324/9780429324178

De Paola, M., & Gioia, F. (2016). Who performs better under time pressure? Results from a field experiment. *Journal of Economic Psychology, 53*, 37–53. https://doi.org/10.1016/j.joep.2015.12.002

Deneen, C. (2020). *Assessment considerations in moving from closed-book to open-book exams*. Melbourne CSHE. https://melbourne-cshe.unimelb.edu.au/__data/assets/pdf_file/0010/3341944/closed-book-to-open-book-exam_final.pdf. Accessed 11/27/2023

Deng, F., Gluckstein, J. A., & Larsen, D. P. (2015). Student-directed retrieval practice is a predictor of medical licensing examination performance. *Perspectives on Medical Education, 4*(6), 308–313. https://doi.org/10.1007/S40037-015-0220-X

DeWitt, S. W., Patterson, N., Blankenship, W., Blevins, B., DiCamillo, L., Gerwin, D., Gradwell, J. M., Gunn, J., Maddox, L., Salinas, C., Saye, J., Stoddard, J., & Sullivan, C. C. (2013). The lower-order expectations of high-stakes tests: A four-state analysis of social studies standards and test alignment. *Theory & Research in Social Education, 41*(3), 382–427. https://doi.org/10.1080/00933104.2013.787031

Durning, S. J., Dong, T., Ratcliffe, T., Schuwirth, L., Artino, A. R., Boulet, J. R., & Eva, K. (2016). Comparing open-book and closed-book examinations: A systematic review. *Academic Medicine, 91*(4), 583–599. https://doi.org/10.1097/ACM.0000000000000977

Ecclestone, K. (1999). Empowerng or ensnaring?: The implications of outcome-based assessment in higher education. *Higher Education Quarterly, 53*(1), 29–48. https://doi.org/10.1111/1468-2273.00111

Ellis, C., van Haeringen, K., Harper, R., Bretag, T., Zucker, I., McBride, S., Rozenberg, P., Newton, P., & Saddiqui, S. (2020). Does authentic assessment assure academic integrity? Evidence from contract cheating data. *Higher Education Research & Development, 39*(3), 454–469. https://doi.org/10.1080/07294360.2019.1680956

Elton, L. R. B. (1987). *Teaching in Higher Education: Appraisal and Training*. London: Kogan Page.

Entwistle, N. J., & Entwistle, A. (1991). Contrasting forms of understanding for degree examinations: The student experience and its implications. *Higher Education, 22*(3), 205–227. https://doi.org/10.1007/BF00132288

Eweda, G., Bukhary, Z. A., & Hamed, O. (2020). Quality assurance of test blueprinting. *Journal of Professional Nursing, 36*(3), 166–170. https://doi.org/10.1016/j.profnurs.2019.09.001

Fang, J., Brown, G. T. L., & Hamilton, R. (2023). Changes in Chinese students' academic emotions after examinations: Pride in success, shame in failure, and self-loathing in comparison. *British Journal of Educational Psychology, 93*(1), 245–261. https://doi.org/10.1111/bjep.12552

Fawaz, Y., & Lee, J. (2022). Rank comparisons amongst teenagers and suicidal ideation. *Economics & Human Biology, 44*, 101093. https://doi.org/10.1016/j.ehb.2021.101093

Fejes, I., Ábrahám, G., & Légrády, P. (2020). The effect of an exam period as a stress situation on baroreflex sensitivity among healthy university students. *Blood Pressure, 29*(3), 175–181. https://doi.org/10.1080/08037051.2019.1710108

Fernández-Castillo, A., & Caurcel, M. J. (2019). Self-esteem, hours of sleep and state-anxiety before academic tests. *Revista Argentina de Clinica Psicologica*, *28*(4), 348–355. https://doi.org/10.24205/03276716.2017.1039

Fernández-Castillo, A., & Caurcel, M. J. (2015). State test-anxiety, selective attention and concentration in university students. *International Journal of Psychology, 50*(4), 265–271. https://doi.org/10.1002/ijop.12092

Fook, C. Y., & Sidhu, G. K. (2014). Assessment practices in higher education in United States. *Procedia - Social and Behavoiral Sceinces, 123*, 299–306.

Franke, M. (2018). Final exam weighting as part of course design. *Teaching and Learning Inquiry*, *6*(1), 91–103. https://doi.org/10.20343/teachlearninqu.6.1.9

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher 18*(9), 27–32. https://www.jstor.org/stable/1176716

Gibbs, G., & Lucas, L. (1997). Coursework assessment, class size and student performance: 1984–94. *Journal of Further and Higher Education, 21*(2), 183–192. https://doi.org/10.1080/0309877970210204

Gibbs, G. (1992). *Improving the quality of student learning*. Technical and Educational Services Ltd.

Gijbels, D., & Dochy, F. (2006). Students' assessment preferences and approaches to learning: Can formative assessment make a difference? *Educational Studies, 32*(4), 399–409. https://doi.org/10.1080/03055690600850354

Gliatto, P., Leitman, I. M., & Muller, D. (2016). Scylla and Charybdis: The MCAT, USMLE, and degrees of freedom in undergraduate medical education. *Academic Medicine, 91*(11), 1498–1500. https://doi.org/10.1097/ACM.0000000000001247

Gómez, R. L., & Suárez, A. M. (2021). Extending impact beyond the community: Protocol for a scoping review of evidence of the impact of communities of practice on teaching and learning in higher education. *International Journal of Educational Research Open, 2*(2), 10048. https://doi.org/10.1016/j.ijedro.2021.100048

Gorgodze, S., & Chakhaia, L. (2021). The uses and misuses of centralised high stakes examinations-Assessment policy and practice in Georgia. *Assessment in Education: Principles, Policy & Practice, 28*(3), 322–342. https://doi.org/10.1080/0969594X.2021.1900775

Greene, E. B. (1931). The retention of information learned in college courses. *The Journal of Educational Research, 24*(4), 262–273. https://doi.org/10.1080/00220671.1931.10880208

Haertel, E. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–100). American Council on Education and Praeger.

Hamzah, F., Mat, K. C., Bhagat, V., & Mahyiddin, N. (2018). Test anxiety and its impact on first year university students and the over view of mind and body intervention to enhance coping skills in facing exams. *Research Journal of Pharmacy and Technology, 11*(6), 2220–2228. https://doi.org/10.5958/0974-360X.2018.00411.0

Harland, T., McLean, A., Wass, R., Miller, E., & Sim, K. N. (2015). An assessment arms race and its fallout: High-stakes grading and the case for slow scholarship. *Assessment and Evaluation in Higher Education, 40*(4), 528–541. https://doi.org/10.1080/02602938.2014.931927

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

Hartog, P., & Rhodes, E. C. (1936). *An examination of examinations*. Macmmillan.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19*(1), 126–134. https://doi.org/10.3758/s13423-011-0181-y

Heeneman, S., de Jong, L. H., Dawson, L. J., Wilkinson, T. J., Ryan, A., Tait, G. R., Rice, N., Torre, D., Freeman, A., & van der Vleuten, C. P. M. (2021). Ottawa 2020 consensus statement for programmatic assessment – 1. Agreement on the Principles. *Medical Teacher, 43*(10), 1139–1148. https://doi.org/10.1080/0142159X.2021.1957088

Hegarty-Kelly, E., & Mooney, D. A. (2021). Analysis of an automatic grading system within first year Computer Science programming modules. *Computing Education Practice, 2021*, 17–20. https://doi.org/10.1145/3437914.3437973

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47–77. https://doi.org/10.3102/00346543058001047

Henderson, M., Ajjawi, R., Boud, D., & Molloy, E. (Eds.). (2020). *The impact of feedback in higher education: Improving assessment outcomes for learners*. Palgrave Macmillan.

Hill, G., Mason, J., & Dunn, A. (2021). Contract cheating: An increasing challenge for global academic community arising from COVID-19. *Research and Practice in Technology Enhanced Learning, 16*(1), 24. https://doi.org/10.1186/s41039-021-00166-8

Högberg, B., & Horn, D. (2022). National high-stakes testing, gender, and school stress in Europe: A difference-in-differences analysis. *European Sociological Review, 38*(6), 975–987. https://doi.org/10.1093/esr/jcac009

Jackson, M., Khavenson, T., & Chirkina, T. (2020). Raising the stakes: Inequality and testing in the Russian education system. *Social Forces, 98*(4), 1613–1635. https://doi.org/10.1093/sf/soz113

Jamil, M. G., Alam, N., Radclyffe-Thomas, N., Islam, M. A., Moniruzzaman Mollah, A. K. M., & Rasel, A. A. (2021). Real world learning and the internationalisation of higher education: Approaches to making learning real for global communities. In D. A. Morley & M. G. Jamil (Eds.), *Applied Pedagogies for Higher Education* (pp. 107–132). Springer International Publishing. https://doi.org/10.1007/978-3-030-46951-1_6

Jones, H., Black, B., Green, J., Langton, P., Rutherford, S., Scott, J., & Brown, S. (2015). Indications of knowledge retention in the transition to higher education. *Journal of Biological Education, 49*(3), 261–273. https://doi.org/10.1080/00219266.2014.926960

Jones, E., Priestley, M., Brewster, L., Wilbraham, S. J., Hughes, G., & Spanner, L. (2021). Student wellbeing and assessment in higher education: The balancing act. *Assessment and Evaluation in Higher Education, 46*(3), 438–450. https://doi.org/10.1080/02602938.2020.1782344

Jones, M. G., Jones, B. D., & Hargrove, T. Y. (2003). *The unintended consequences of high-stakes testing*. Rowman & Littlefield Pubishers.

Karami, H. (2013). An investigation of the gender differential performance on a high-stakes language proficiency test in Iran. *Asia Pacific Education Review, 14*(3), 435–444. https://doi.org/10.1007/s12564-013-9272-y

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Kellaghan, T., & Greaney, V. (2019). *Public examinations examined*. World Bank. https://doi.org/10.1596/978-1-4648-1418-1

Kickert, R., Meeuwisse, M., Stegers-Jager, K. M., Prinzie, P., & Arends, L. R. (2022). Curricular fit perspective on motivation in higher education. *Higher Education, 83*(4), 729–745. https://doi.org/10.1007/s10734-021-00699-3

Klenowski, V. (2009). Australian Indigenous students: Addressing equity issues in assessment. *Teaching Education, 20*(1), 77–93.

Klenowski, V. (2016). Fairer assessment for indigenous students: An Australian perspective. In S. Scott, D. E. Scott, & C. F. Webber (Eds.), *Leadership of assessment, inclusion, and learning* (Vol. 3, pp. 273–285). Springer International Publishing. https://doi.org/10.1007/978-3-319-23347-5_11

Knight, P. T. (2002). Summative assessment in higher education: Practices in disarray. *Studies in Higher Education, 27*(3), 275–286. https://doi.org/10.1080/03075070220000662

Knight, P. T., & Yorke, M. (2003). *Assessment, learning and employability*. Society for Research into Education & Open University Press.

Kuhbandner, C., Aslan, A., Emmerdinger, K., & Murayama, K. (2016). Providing extrinsic reward for test performance undermines long-term memory acquisition. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00079

Kumar, V. S., & Boulanger, D. (2021). Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education, 31*(3), 538–584. https://doi.org/10.1007/s40593-020-00211-5

Kwon, S. K., Lee, M., & Shin, D. (2017). Educational assessment in the Republic of Korea: Lights and shadows of high-stake exam-based education system. *Assessment in Education: Principles, Policy & Practice, 24*(1), 60–77. https://doi.org/10.1080/0969594X.2015.1074540

Lancaster, T., & Clarke, R. (2017). Rethinking assessment by examination in the age of contract cheating. *Plagiarism across Europe and Beyond 2017—Conference Proceedings*, 215–228.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine, 6*(7), e1000100. https://doi.org/10.1371/journal.pmed.1000100

Lotz, C., & Sparfeldt, J. R. (2017). Does test anxiety increase as the exam draws near? – Students' state test anxiety recorded over the course of one semester. *Personality and Individual Differences, 104*, 397–400. https://doi.org/10.1016/j.paid.2016.08.032

Maes, M., Van Der Planken, M., Van Gastel, A., Bruyland, K., Van Hunsel, F., Neels, H., Hendriks, D., Wauters, A., Demedts, P., Janca, A., & Scharpé, S. (1998). Influence of academic examination stress on hematological measurements in subjectively healthy volunteers. *Psychiatry Research, 80*(3), 201–212. https://doi.org/10.1016/S0165-1781(98)00059-6

Marchant, G. J., & Paulson, S. E. (2005). The relationship of high school graduation exams to graduation rates and SAT scores. *Education Policy Analysis Archives*, *13*(6). https://files.eric.ed.gov/fulltext/EJ846516.pdf. Accessed 11/27/2023

Marton, F., & Säljö, R. (1997). Approaches to learning. In F. Marton, D. Hounsell, & N. Entwistle (Eds.), *The experience of learning. Implications for teaching and studying in higher education.* (pp. 39–59). Scottish Academic Press. http://www.docs.hss.ed.ac.uk/iad/Learning_teaching/Academic_teaching/Resources/Experience_of_learning/EoLChapter3.pdf. Accessed 11/27/2023

Mason, E. J. (2007). Measurement issues in high stakes testing: Validity and reliability. *Journal of Applied School Psychology, 23*(2), 27–46. https://doi.org/10.1300/J370v23n02_03

McCabe, D. L. (2005). Cheating among college and university students: A North American perspective. *International Journal for Educational Integrity*, *1*(1). https://doi.org/10.21913/IJEI.v1i1.14

McConnell, M. M., St-Onge, C., & Young, M. E. (2015). The benefits of testing for learning on later performance. *Advances in Health Sciences Education, 20*(2), 305–320. https://doi.org/10.1007/s10459-014-9529-1

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4–5), 494–513. https://doi.org/10.1080/09541440701326154

Meeks, L. M., Plegue, M., Swenor, B. K., Moreland, C. J., Jain, S., Grabowski, C. J., Westervelt, M., Case, B., Eidtson, W. H., Patwari, R., Angoff, N. R., LeConche, J., Temple, B. M., Poullos, P., Sanchez-Guzman, M., Coates, C., Low, C., Henderson, M. C., Purkiss, J., & Kim, M. H. (2022). The performance and trajectory of medical students with disabilities: Results from a multisite, multicohort study. *Academic Medicine, 97*(3), 389–397. https://doi.org/10.1097/ACM.0000000000004510

Mehrazmay, R., Ghonsooly, B., & De La Torre, J. (2021). Detecting differential item functioning using cognitive diagnosis models: Applications of the Wald test and likelihood ratio test in a university entrance examination. *Applied Measurement in Education, 34*(4), 262–284. https://doi.org/10.1080/08957347.2021.1987906

Messick, S. (1992). Validity of test interpretation and use. In *Encyclopedia of educational research* (6th ed., pp. 1487–1495). Macmillan. https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.1990.tb01343.x. Accessed 11/27/2023

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology, 68*(1), 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Miller, K., Lukoff, B., King, G., & Mazur, E. (2018). Use of a social annotation platform for pre-class reading assignments in a flipped introductory physics class. *Frontiers in Education, 3*, 8. https://doi.org/10.3389/feduc.2018.00008

Monrad, S. U., Wolff, M., Kurtz, J., Deiorio, N. M., Sabo, R., Stringer, J. K., & Santen, S. A. (2021). What is the association between student well-being and high-stakes examination scores? *Medical Education, 55*(7), 872–877. https://doi.org/10.1111/medu.14460

Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education*, *9*(3). https://doi.org/10.1002/rev3.3292

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology, 18*(1), 143. https://doi.org/10.1186/s12874-018-0611-x

Mutereko, S. (2018). Marketisation, managerialism and high-stake testing: A tale teachers' views on national assessments in South Africa. *International Journal of Educational Management, 32*(4), 568–579. https://doi.org/10.1108/IJEM-04-2017-0096

Nieminen, J. H., & Tuohilampi, L. (2020). 'Finally studying for myself' – Examining student agency in summative and formative self-assessment models. *Assessment & Evaluation in Higher Education, 45*(7), 1031–1045. https://doi.org/10.1080/02602938.2020.1720595

Nieminen, J. H. (2022). Assessment for inclusion: Rethinking inclusive assessment in higher education. *Teaching in Higher Education*, 1–19. https://doi.org/10.1080/13562517.2021.2021395

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2019). Gender-based differential prediction by curriculum samples for college admissions. *Educational Measurement: Issues and Practice, 38*(3), 33–45. https://doi.org/10.1111/emip.12266

Pascoe, M. C., Hetrick, S. E., & Parker, A. G. (2020). The impact of stress on students in secondary school and higher education. *International Journal of Adolescence and Youth, 25*(1), 104–112. https://doi.org/10.1080/02673843.2019.1596823

Peh, L. L. C., Cerimagic, S., & Conejos, S. (2021). Challenges of running online exams and preventing academic dishonesty during the Covid-19 pandemic. *Journal of Learning Development in Higher Education*, *22*. https://doi.org/10.47408/jldhe.vi22.830

Potaka, E., & Huang, C. (2015). Pens for hire: How students cheat, and how they get away with it. *SBS*. https://www.sbs.com.au/news/the-feed/article/pens-for-hire-how-students-cheat-and-how-they-get-away-with-it/5v3erlpij. Accessed 11/27/2023

Preston, J. P., & Claypool, T. R. (2021). Analyzing assessment practices for indigenous students. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.679972

Raman, R., Sairam, B., Veena, G., Vachharajani, H., & Nedungadi, P. (2021). Adoption of online proctored examinations by university students during COVID-19: Innovation diffusion study. *Education and Information Technologies, 26*(6), 7339–7358. https://doi.org/10.1007/s10639-021-10581-5

Ramsden, P. (1992). *Learning to teach in higher education*. Routledge.

Ramsden, P. (1997). The context of learning in academic departments. In F. Marton, D. Hounsell, & N. Entwistle (Eds.), *The experience of learning. Implications for teaching and studying in higher education* (pp. 198–217). Scottish Academic Press. http://www.docs.hss.ed.ac.uk/iad/Learning_teaching/Academic_teaching/Resources/Experience_of_learning/EoLChapter13.pdf. Accessed 11/27/2023

Rask, K., & Tiefenthaler, J. (2008). The role of grade sensitivity in explaining the gender imbalance in undergraduate economics. *Economics of Education Review, 27*(6), 676–687. https://doi.org/10.1016/j.econedurev.2007.09.010

Rawlusyk, P. E. (2018). Assessment in higher education and student learning. *Journal of Instructional Pedagogies*, *21*, 1–34. https://files.eric.ed.gov/fulltext/EJ1194243.pdf. Accessed 11/27/2023

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review, 25*(4), 523–548. https://doi.org/10.1007/s10648-013-9240-4

Reedy, A., Pfitzner, D., Rook, L., & Ellis, L. (2021). Responding to the COVID-19 emergency: Student and academic staff perceptions of academic integrity in the transition to online exams at three Australian universities. *International Journal for Educational Integrity*, *17*(1). https://doi.org/10.1007/s40979-021-00075-9

Richardson, J. T. E. (2015a). Coursework versus examinations in end-of-module assessment: A literature review. *Assessment and Evaluation in Higher Education, 40*(3), 439–455. https://doi.org/10.1080/02602938.2014.919628

Richardson, J. T. E. (2015b). The under-attainment of ethnic minority students in UK higher education: What we know and what we don't know. *Journal of Further and Higher Education, 39*(2), 278–291. https://doi.org/10.1080/0309877X.2013.858680

Roediger, H. L., & Butler, A. C. (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., & Karpicke, J. D. (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

Roos, A. L., Goetz, T., Voracek, M., Krannich, M., Bieg, M., Jarrell, A., & Pekrun, R. (2021). Test anxiety and physiological arousal: A systematic review and meta-analysis. *Educational Psychology Review, 33*(2), 579–618. https://doi.org/10.1007/s10648-020-09543-z

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Ryan, T., French, S., & Kennedy, G. (2021). Beyond the Iron Triangle: Improving the quality of teaching and learning at scale. *Studies in Higher Education, 46*(7), 1383–1394. https://doi.org/10.1080/03075079.2019.1679763

Sahlberg, P., & Hargreaves, A. (2011). *Finnish lessons: What can the world learn from educational change in Finland?* Teachers College Press.

Salehi, S., Cotner, S., Azarin, S. M., Carlson, E. E., Driessen, M., Ferry, V. E., Harcombe, W., McGaugh, S., Wassenberg, D., Yonas, A., & Ballen, C. J. (2019). Gender performance gaps across different assessment methods and the underlying mechanisms: The case of incoming preparation and test anxiety. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00107

Sambell, K., McDowell, L., & Brown, S. (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

Santovena-Casal, S. (2019). Effects of continuous assessment on the academic performance of future teachers. *Croatian Journal of Education*, *21*(3), 777–822. https://doi.org/10.15516/cje.v21i3.3013

Shaw, S., Crisp, V., & Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy & Practice, 19*(2), 159–176. https://doi.org/10.1080/0969594X.2011.563356

ShayesteFar, P. (2020). A model of interplay between student English achievement and the joint affective factors in a high-stakes test change context: Model construction and validity. *Educational Assessment Evaluation and Accountability, 32*(3), 335–371. https://doi.org/10.1007/s11092-020-09326-8

Shean, M. (2019). Don't calm down! Exam stress may not be fun but it can help you get better marks. *The Conversation*. https://theconversation.com/dont-calm-down-exam-stress-may-not-be-fun-but-it-can-help-you-get-better-marks-124517. Accessed 11/27/2023

Sheard, J., & Dick, M. (2003). Influences on cheating practice of graduate students in it courses: What are the factors? *Proceedings of the Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (ITiSCE)*, *8*(September), 45–49. https://doi.org/10.1145/961290.961527

Simonite, V. (2003). The impact of coursework on degree classifications and the performance of individual students. *Assessment & Evaluation in Higher Education, 28*(5), 459–470. https://doi.org/10.1080/02602930301675

Slavin, S. J., Schindler, D. L., & Chibnall, J. T. (2014). Medical student mental health 3.0: Improving student wellness through curricular changes. *Academic Medicine*, *89*(4), 573–577. https://doi.org/10.1097/ACM.0000000000000166

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20*(5), 8–11. https://doi.org/10.3102/0013189X020005008

Smith, M. L., & Fey, P. (2000). Validity and accountability in high-stakes testing. *Journal of Teacher Education, 51*(5), 334–344. https://doi.org/10.1177/0022487100051005002

Sommer, M., & Arendasy, M. E. (2015). Further evidence for the deficit account of the test anxiety–test performance relationship from a high-stakes admission testing setting. *Intelligence, 53*, 72–80. https://doi.org/10.1016/j.intell.2015.08.007

Søndergaard, H., & Mulder, R. A. (2012). Collaborative learning through formative peer review: pedagogy, programs and potential. *Computer Science Education, 22*(4), 343–367. https://doi.org/10.1080/08993408.2012.728041

Song, X. (2016). Fairness in educational assessment in China: Historical practices and contemporary challenges. In S. Scott, D. E. Scott, & C. F. Webber (Eds.), *Assessment in education* (Vol. 2, pp. 67–89). Springer International Publishing. https://doi.org/10.1007/978-3-319-23398-7_4

Stenlund, T., Lyrén, P.-E., & Eklöf, H. (2018). The successful test taker: Exploring test-taking behavior profiles through cluster analysis. *European Journal of Psychology of Education, 33*(2), 403–417. https://doi.org/10.1007/s10212-017-0332-2

Sternberg, R. J. (1997). *Successful intelligence: How practical and creative intelligence determine success in life*. Plume.

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research, 51*(2), 161–179. https://doi.org/10.1080/00131880902891305

Stopar, A., & Ilc, G. (2017). Reading for a test: The effect of high-stakes exams on reading strategies. *Porta Linguarum, Monográfico II:* 103–115. https://doi.org/10.30827/Digibug.54115

Surgenor, P. W. G. (2013). Measuring up: Comparing first year students' and tutors' expectations of assessment. *Assessment and Evaluation in Higher Education, 38*(3), 288–302. https://doi.org/10.1080/02602938.2011.630976

Sutherland-Smith, W. (2008). *Plagiarism, the internet, and student learning: Improving academic integrity.* Routledge.

Tai, J., Ajjawi, R., Bearman, M., Boud, D., Dawson, P., & Jorre de St Jorre, T. (2022). Assessment for inclusion: Rethinking contemporary strategies in assessment design. *Higher Education Research & Development*, 1–15. https://doi.org/10.1080/07294360.2022.2057451

Theobald, M., Breitwieser, J., & Brod, G. (2022). Test anxiety does not predict exam performance when knowledge is controlled for: Strong evidence against the interference hypothesis of test anxiety. *Psychological Science, 33*(12), 2073–2083. https://doi.org/10.1177/09567976221119391

Trotter, E. (2006). Student perceptions of continuous summative assessment. *Assessment and Evaluation in Higher Education, 31*(5), 505–521. https://doi.org/10.1080/02602930600679506

Trumbull, E., & Nelson-Barber, S. (2019). The ongoing quest for culturally-responsive assessment for indigenous students in the U.S. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00040

Uy, C., Manalo, R. A., & Cabauatan, R. R. (2015). Factors affecting university entrants' performance in high-stakes tests: A multiple regression analysis. *Asia Pacific Education Review, 16*(4), 591–601. https://doi.org/10.1007/s12564-015-9395-4

Van Bergen, P., & Lane, R. (2014). Exams might be stressful, but they improve learning. *The Conversation*. https://theconversation.com/exams-might-be-stressful-but-they-improve-learning-35614. Accessed 11/27/2023

Vaughan, N. (2014). Student engagement and blended learning: Making the assessment connection. *Education Sciences, 4*(4), 247–264. https://doi.org/10.3390/educsci4040247

Verdake, H., Mulhern, T. D., Lodge, J., Elliott, K., Cropper, S., Rubinstein, B., Horton, A., Elliott, C., Espinosa, A., Dooley, L., Frankland, S., Mulder, R., & Livett, M. (2017). *Misconceptions as a trigger for enhancing student learning in higher education*. The University of Melbourne.

Villarroel, V., Boud, D., Bloxham, S., Bruna, D., & Bruna, C. (2019). Using principles of authentic assessment to redesign written examinations and tests. *Innovations in Education and Teaching International*, 1–12. https://doi.org/10.1080/14703297.2018.1564882

Villaroel, V., Boud, D., Bloxham, S., Bruna, D., & Bruna, C. (2020). Using principles of authentic assessment to redesign written examinations and tests. *Innovations in Education and Teaching International, 57*(1), 38–49. https://doi.org/10.1080/14703297.2018.1564882

Vogel, S., & Schwabe, L. (2016). Learning and memory under stress: Implications for the classroom. *Npj Science of Learning, 1*(1), 16011. https://doi.org/10.1038/npjscilearn.2016.11

Von Der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders, 227*, 483–493. https://doi.org/10.1016/j.jad.2017.11.048

Wang, Z. L., & Brown, G. T. (2014). Hong Kong tertiary students' conceptions of assessment of academic ability. *Higher Education Research & Development, 33*(5), 1063–1077. https://doi.org/10.1080/07294360.2014.890565

Wang, J., Li, Q., & Luo, Y. (2022). Physics identity of chinese students before and after Gaokao: The effect of high-stake testing. *Research in Science Education, 52*(2), 675–689. https://doi.org/10.1007/s11165-020-09978-y

Wass, R., Harland, T., McLean, A., Miller, E., & Sim, K. N. (2015). 'Will press lever for food': Behavioural conditioning of students through frequent high-stakes assessment. *Higher Education Research and Development, 34*(6), 1324–1326. https://doi.org/10.1080/07294360.2015.1052351

Weekes, N., Lewis, R., Patel, F., Garrison-Jakel, J., Berger, D. E., & Lupien, S. J. (2006). Examination stress as an ecological inducer of cortisol and psychological responses to stress in undergraduate students. *Stress, 9*(4), 199–206. https://doi.org/10.1080/10253890601029751

Williams, P. (2008). Assessing context-based learning: Not only rigorous but also relevant. *Assessment & Evaluation in Higher Education, 33*(4), 395–408. https://doi.org/10.1080/02602930701562890

Williams, P. (2014). Squaring the circle: A new alternative to alternative-assessment. *Teaching in Higher Education, 19*(5), 565–577. https://doi.org/10.1080/13562517.2014.882894

Winstone, N. E., & Carless, D. (2020). *Designing effective feedback processes in higher education: A learning-focused approach*. Routledge, Taylor & Francis Group.

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *The Journal of General Education, 58*(3), 152–166.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*(3), 227–242. https://doi.org/10.1207/s15324818ame0803_3

Wong, H. M., Kwek, D., & Tan, K. (2020). Changing assessments and the examination culture in Singapore: A review and analysis of Singapore's assessment policies. *Asia Pacific Journal of Education, 40*(4), 433–457. https://doi.org/10.1080/02188791.2020.1838886

Woodfield, R., Earl-Novell, S., & Solomon, L. (2005). Gender and mode of assessment at university: Should we assume female students are better suited to coursework and males to unseen examinations?1. *Assessment & Evaluation in Higher Education, 30*(1), 35–50. https://doi.org/10.1080/0260293042003243887

Zhan, Y., & Andrews, S. (2014). Washback effects from a high-stakes examination on out-of-class English learning: Insights from possible self theories. *Assessment in Education: Principles, Policy & Practice, 21*(1), 71–89. https://doi.org/10.1080/0969594X.2012.757546

Zhang, Z., Su, H., Peng, Q., Yang, Q., & Cheng, X. (2011). Exam anxiety induces significant blood pressure and heart rate increase in college students. *Clinical and Experimental Hypertension, 33*(5), 281–286. https://doi.org/10.3109/10641963.2010.531850

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.