Received: 27 January 2025 | Accepted: 3 December 2025

DOI: 10.1002/rev3.70126

REVIEW AND SYNTHESIS

Review of Education BERA

What do I have to do to get the grade? How examination standards interact with learning, teaching and the curriculum

Michelle Meadows[1] | **Stuart Cadwallader**[1] | **Lena Gray**[2] | **Jo-Anne Baird**[1]

[1]Department of Education, University of Oxford, Oxford, UK

[2]Independent Consultant, Glasgow, UK

Correspondence
Michelle Meadows, Department of Education, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, UK.
Email: michelle.meadows@education.ox.ac.uk

Funding information
Qualifications Wales, UK, Grant/Award Number: QW222313

Abstract

Communicating national qualification standards clearly to learners and their teachers is crucial to raising standards. If people do not know what they must do to get the grade, then the qualification is providing poor information about what is considered valuable learning. Assessment scores (and grades) need to convey meaning about learners' knowledge and skills. Yet score meaning is obscure, which leads to frustration on the part of teachers and learners. How standards are conceived, and set, interacts with score meaning. In this article, we outline how the current system of setting standards for GCSEs and A-levels in England, Wales and Northern Ireland—attainment-referencing—affects the clarity of grading standards. Attainment-referencing offers no single artefact to represent standards, as it is the product of an amalgam of evidence. This makes it difficult for learners and their teachers to derive score meaning from their assessment outcomes. We contrast this with other methods for setting and maintaining standards that have been suggested as better alternatives: norm- and criterion-referencing. Both, in their own ways, purport to offer greater clarity. However, norm-referencing does not relate directly to curriculum standards, but to learners' ranks with respect to a population. Criterion-referencing relates directly to the curriculum but operates better in theory than in practise—criteria can never be specified

unproblematically because of the limits of language. Within the context of how standards are currently set and maintained, attainment-referencing, we suggest ways of better explaining the curriculum-related meaning of national qualification grades.

KEYWORDS
attainment-referencing, criterion-referencing, norm-referencing, score meaning

Context and implications

Rationale for this study and why the new findings matter: Transparency of the requirements for the awarding of school exam grades is essential for supporting teaching and learning. Here, for the first time in the literature, we outline how approaches to standards—norm-referencing, criterion-referencing and attainment-referencing—impact grading transparency. Differences between school and higher education contexts are considered.

Implications for researchers, practitioners and policy makers: Periodically, public debates regarding the best way to set standards arise. Our analysis shows that whilst criterion-referencing and norm-referencing are transparent in some respects, neither delivers what teachers, students and other users of the exam results need to understand grading outcomes. Attainment-referencing takes into account aspects of the assessment context in which students perform. It is the system used for GCSEs and A-levels and has clear advantages in that context, but its reliance on multiple sources of evidence means it too fails to deliver complete transparency of grading requirements. Ultimately, perfect transparency of standards is unachievable in school or higher education assessments.

INTRODUCTION

Educational assessments take place in different contexts across the world; they are ubiquitous in higher education institutions, schools and workplaces. Many assessment concepts are transposable across these different contexts, such as validity, reliability and fairness, and there are common assessment formats. Each context generates policies, tools and regulations for addressing the common problems that assessment administrators face, such as maintaining standards, aggregating across assessment components and establishing consistent marking. Due to the varied social contexts in which assessments are conducted, assessment systems have differing priorities, such as certification or selection. They also differ in the features that are socially accepted and aspects that would be considered scandalous if they were to go wrong (Baird et al., 2018).

This article, set within the Special Issue on Transparency in Assessment, which focuses mainly on a higher education context, seeks to explain how different ways of setting standards affect the transparency of grading outcomes in school examinations in England, Wales and Northern Ireland. We note that the assessment literature in higher education is highly active, as is the broader assessment literature, but these literatures do not often speak to

each other. In part therefore, we aim to make connections between the assessment literature in higher education and the assessment literature on school qualifications. As such, this paper addresses both audiences, and so we must explain some of the differences in context before turning to the cases that form our focus. Ultimately, we draw conclusions for both contexts.

In higher education, marking and grading are typically conflated, whereas they are usually separate processes in school qualifications. Assessment-specific marking schemes (rubrics) are generated for school qualifications, whereas generic grading criteria are typically used in higher education. A standard-setting process is required to turn marks (scores) into grades in school qualifications, but scores have the same grade worthiness across assessments in higher education: a score of 70 translates to the same grade each time the assessment is administered. Where materials are produced to communicate standards, they tend to be for learners in higher education, whereas they are often designed for teachers in school qualifications.

Regulatory bodies and their impacts on the design of assessments differ considerably across these contexts. Most assessments in higher education are approved within the institution, whereas they are highly specified by governmental bodies for the school qualifications featured in this article. Exam boards that operate the qualifications have expertise in assessment; that is their primary organisational purpose. In higher education, the purpose is broadly educational, and there may be little assessment expertise, even if there is a great deal of practitioner experience. National school qualifications are under more public and political scrutiny for the ways in which assessments are operated, though academic standards in higher education have come under mounting scrutiny (Rust & O'Donovan, 2025).

This article primarily focuses on outlining the implications of three different approaches to standard setting in school qualifications and discusses the extent to which they can deliver transparency in this context. However, these conceptions of assessment standards also offer lessons for higher education. Identifying the approach to standard setting that is in use, assumed, or being considered is helpful for making different views and practises explicit and for unpacking their implications, promises and shortfalls.

THE DESIRE FOR TRANSPARENT EDUCATIONAL GOALS FOR SCHOOL SYSTEMS

National qualifications, such as GCSEs and A-levels in England, Wales and Northern Ireland, set out goals for the education system about what learners are supposed to know and be able to do. The assumption is that setting transparent goals via qualifications will promote higher educational standards, as teachers will align their instruction with the desired outcomes and learners will understand what is expected, allowing for more targeted learning. Each GCSE and A-level qualification is underpinned by a specification (syllabus), outlining the content to be taught and the method through which it will be assessed. It is through the assessment itself, however, that the core knowledge and skills at the heart of a qualification (the construct) are made most transparent. As Wiliam (2010) argued, assessments are how the curriculum is operationalised, they make the construct concrete.

Cizek (2016, 2020) used the term ‘score meaning’ to describe the extent to which a mark or grade may be interpreted in terms of what the learner knows and can do in relation to the construct. This notion is central to assessment validity, and without it learning goals cannot be communicated with clarity. In the context of GCSEs and A-levels meaning is generally drawn from the learner's grade rather than their ‘score’. Grades are determined by aggregating multiple weighted assessment scores rather than by performance on a single test. Grades therefore reflect achievement across the full range of assessments a learner

completes to obtain the qualification. This will encompass a range of information from different subject-specific forms of assessment (such as exams and coursework) that are designed to cover the entirety of the specified content. Arguably, then, a more precise term in the context of GCSEs, A-levels and similar qualifications, might be 'grade meaning' rather than 'score meaning'. Though this is an important distinction, this paper will use the term of 'score meaning' to refer to the information that a given assessment outcome provides about the learners' knowledge and skills as it is used more commonly in the literature.

For GCSEs and A-levels, beyond the provision of specifications, there are two main approaches to explicating constructs and the meaning of scores derived from the assessment process. First, the knowledge and skills required for specific grades may be set out in grade descriptors (see, e.g., JCQ, 2021). These descriptions are designed to give teachers and learners an indication of the likely level of performance that would be expected for a given grade. They seek to make examiners' implicit understanding of performance standards explicit (Greatorex, 2002; Sadler, 1987). However, grade descriptors are usually difficult to directly connect with learning activities because they are broad, encompassing a wide range of possible learner performances by describing them generically (Cresswell, 1987; Greatorex, 2005). An example description of grade 5 performance in GCSE Mathematics is provided below.

To achieve grade 5, candidates will be able to:

- Perform routine single- and multi-step procedures effectively by recalling, applying and interpreting notation, terminology, facts, definitions and formulae.
- Interpret and communicate information effectively.
- Make deductions, inferences and draw conclusions.
- Construct chains of reasoning, including arguments.
- Generate strategies to solve mathematical and non-mathematical problems by translating them into mathematical processes, realising connections between different parts of mathematics.
- Interpret results in the context of the given problem.
- Evaluate methods and results.

Ofqual (2023, para. 2)

Second, exam boards publish question papers and mark schemes, alongside examiner reports and boundary marks (cut-scores) for each grade. This information more directly links performance to grades, outlining the minimum number of marks required for each grade in a given year for a specific assessment. However, the assessment tasks change each year in ways that are unpredictable. They are designed to be predictable enough for effective preparation, yet not so predictable that they encourage teaching and learning 'to the test' and inauthentic performances that do not reflect learners' genuine, underlying attainment (Au, 2007; Holmes et al., 2020; Popham, 2001). The need to create some unpredictability in assessments means that the link between learning and assessment is also somewhat unpredictable, obscuring score meaning. Knowledge in a specific area of the curriculum may have greater weight in 1 year compared with another, depending on the exact questions that appear in the assessment. For example, although there will always be questions about the periodic table in GCSE Chemistry examinations, each will focus upon different chemical elements or aspects of those elements (bonds, reactivity, atomic number, etc.).

Hence, despite the publication of grade descriptors, question papers, marking schemes and so on, there remains a lack of transparency between the curriculum (the content that society values), the grading standards (the performance required at each grade) and, crucially, teaching and learning. Score meaning is ambiguous, making it difficult for teachers and learners to know, with the levels of precision they desire, what standards are required.

One consequence of this opacity is the oft-repeated question from learners to their teachers: *Can you tell me what I have to do to get the grade?* Understandably, it is frustrating for learners if they do not get a straightforward answer. After all, dedicated teachers and learners are motivated to understand the material so that their endeavours manifest the knowledge and skills that are outlined through the curriculum in the way that is valued in the marking scheme.

There have been calls for alternative approaches to the process of setting standards for GCSEs and A-levels (e.g., Alliance for Workers' Liberty, 2021; Lucas et al., 2020; Rosen, 2017). These demands stem from the lack of clarity about the standards required to achieve grades and a growing suspicion that, under the current system of attainment-referencing, grades may be artificially inflated or suppressed (Benton, 2016; Newton, 2022). Norm-referencing and particularly criterion-referencing are often suggested as alternatives to the current approach. For example, McArthur (2020) argued that criterion-referencing in GCSEs and A-levels would ensure 'that assessment results genuinely relate to the individual student's performance' (para. 9). Whereas Richmond (2021) proposed that GCSE grades be scrapped and that 'each student's overall score as well as their percentile rank' (p. 6) be reported. Leech (2023) provided a review of seven of the numerous recent reports from think tanks, examination boards and commissions, each of which set out different visions for the future of GCSEs and A-levels. Recent widespread interest in the design and delivery of GCSEs and A-levels, is related to their imminent reform in England. One aim of the reforms is to address the socioeconomic gap for educational attainment (Department for Education, 2025).

In the context of higher education, it has been claimed that transparency through explicit assessment criteria is essential to promote equality of opportunity among learners. It is argued that learners from lower socioeconomic or ethnic minority backgrounds will benefit most from clarity about the standards required to achieve grades (Balloo et al., 2018). Nonetheless, scholars argue it is important to recognise the dynamic, tacit nature of standards which makes it impossible to codify them fully (Ajjawi et al., 2021; Hudson et al., 2015; Sadler, 2009). Active engagement by learners, teachers and examiners with assessment criteria through dialogue and interaction is vital then, rather than relying exclusively on explicit written documentation (Gonsalves & Lin, 2024). Certainly, transparency is seen as an important element of equitable assessment practise in higher education.

Debates around transparency are framed differently for GCSEs and A-levels. While it has long been argued that 'The best defence against inequitable assessment is openness' (Gipps, 1999, p. 385), scholarly focus has been on the pressures of school accountability on teaching and learning (Santori, 2020). It is believed that these pressures have led teachers to overly focus on learners' understanding of assessment rubrics and the practise of past examinations (Daly et al., 2012; Perryman et al., 2011). There is also concern that too much transparency about what is tested could limit learning to lower-order skills, reducing it to memorisation of highly specified performances (Baird et al., 2013).

Given the demands of current school accountability arrangements, many stakeholders' concerns relate to a perceived lack of clarity regarding the specific performances required for learners to achieve grades. For example, in 2024 there was a petition calling for greater transparency in setting grade boundaries following the award of GCSE English language in which there was a significant increase in the minimum number of marks required for each grade compared to previous years (Singh, 2024). As a result, evaluating potential benefits to transparency of alternative approaches to standard setting is timely, especially given that qualification reform is on the horizon.

In this article, we examine the current approach to setting standards in GCSEs and A-levels and explore whether alternatives could offer greater clarity to teachers and learners regarding the expected standard. We explore the meaning of attainment-, norm- and

criterion-referencing, why different approaches are attractive to those seeking transparency, how they might support teachers and learners, and ultimately, what they could deliver in terms of clarity about the performance standards required of learners.

In this way, we contribute to the discussion on how best to enhance the relationship between assessment and learning in high-stakes national qualifications, such as GCSEs and A-levels (Baird et al., 2017). This is important as debates regarding the use of norm-referencing, criterion-referencing or attainment-referencing emerge periodically, yet there is no source in the literature that unpacks the relationship between each of these standard setting approaches and the transparency of educational goals. We conclude the article by arguing that whilst the promise of norm- and criterion-referencing is clarity in what is required of learners to achieve grades, they do not deliver what is currently required from school qualifications in the education system in England, Wales and Northern Ireland. Drawing on literature from both school and higher education contexts, we outline what more could be done to improve transparency in the standards of performance required to achieve grades under attainment-referencing.

METHOD

The novel contribution of this paper is predicated on an application of assessment theory on standards to the practise of designing and operating national assessment systems, focusing on the case of GCSEs and A-levels. A critical integrative review approach was utilised, combining a review of the academic literature with professional expertise gained from the authors' previous senior roles in national assessment systems.

The article builds upon recent projects on conceptualisations of standards in qualifications in Wales (Meadows et al., 2023), public perceptions of standards in Scotland (Hayward et al., 2023) and a comparative project investigating how standards were set in national systems around the world (Baird et al., 2018). The current article is a state-of-the-art review, integrating concepts and theory from sub-fields of the educational assessment literature to analyse prospects for transparency in different approaches to theorising standards.

The Scopus database and Google Scholar were searched, with the latter proving particularly useful for identifying relevant grey literature. The resulting corpus was supplemented by key sources in the field identified by the authors. The review did not aim to be exhaustive, rather the inclusion of papers was based on conceptual relevance. Further, many papers relevant to the topic of standard setting were theoretical or observational rather than empirical. We deliberately engaged with seminal works that have defined and advanced the principal approaches to standard setting. This included foundational texts by Glaser (1963) and Angoff (1974) on criterion- and norm-referencing, Newton (2022) on attainment-referencing, and critical analyses such as that by Torrance (2007) in relation to criterion-referencing. Each conception of standards was situated within its corresponding assessment paradigm, drawing on and extending the theoretical framework of assessment paradigms proposed by Baird and Opposs (2018).

Assessment paradigms

The theoretical framework employed distinguishes different logics of assessment, going beyond discussions of format, operation or contexts of assessment to draw upon different paradigms of assessment (Baird & Opposs, 2018). The paradigms referred to are psychometric-, curriculum- and outcomes-based. Each has different:

- Ways of conceiving of the attribute being assessed.
- Typical instrument formats.
- Purpose of assessment.
- Prioritisation of quality indicators.
- Perspectives on the meaning of standards and which standard setting approaches are necessary.

High stakes tests have been identified operating under each of these paradigms in different countries (Baird et al., 2018).

In outlining the paradigms, it is helpful to think of prototypical cases. Intelligence testing operates under a psychometrics-based paradigm, school qualifications, such as the GCSEs and A-levels, operate in a curriculum-based paradigm and vocational assessments are outcomes-based. The psychometrics-based paradigm largely concerns the measurement of psychological constructs—latent traits that cannot be directly observed. Psychometric statistics are used to reduce measurement error in assessments to gain a more accurate estimation of a person's trait (or, in this context, ability). In contrast, curriculum-based assessments address learners' attainment against the specified curriculum. And in outcomes-based assessments, the attribute of interest is competency, which may be demonstrated through a range of observable performances, often involving a practical skill. We describe how each of the standard setting approaches is embedded in a different assessment paradigm. We then present an integrative argument, summarised in Table 1 at the end of the article, drawing out the pertinent aspects of each paradigm in relation to the three standard-setting approaches and their prospects for improved score meaning.

The case of GCSEs and A-levels

The three forms of standard setting listed above are internationally accepted assessment terminology. To explain the ways and extents to which they elucidate score meaning for teachers and learners, however, we have made the methodological choice to apply the terms to the case of GCSEs and A-levels in England, Wales and Northern Ireland. The literature on the meaning of assessment standards demonstrates that it is culturally bound. What is acceptable in one context is interpreted very differently in others (Isaacs & Gorgen, 2018). Expectations for inter-rater comparability, national standardisation and ultimately score meaning will therefore vary across countries. For example, in some countries, such as France or Ireland, the boundary mark to be awarded a particular grade remains the same every year and there would be an outcry if there were changes (Gauthier, 2018; McManus, 2018). But in England, standard setting processes alter the grade boundary marks slightly each year, to allow for differences in difficulty of the examinations (Taylor & Opposs, 2018). Gauthier (2018, p. 126) also argued that in France, 'policymakers and the public accept a relative ignorance about what the baccalauréat as a whole checks and proves'. Whereas in England, Wales and Northern Ireland, there is lively debate about the strengths and weakness of the qualification system among many stakeholders including politicians, policymakers and some teachers (Leech, 2023).

Professional expertise and positionality

The authors possess extensive professional experience within examination boards and regulatory bodies, having actively contributed to the design, implementation and evaluation of national qualifications including GCSEs, A-levels and vocational and technical qualifications. Our roles

TABLE 1 Unpacking score meaning for different standard setting approaches.

Element	Criterion-referencing	Norm-referencing	Attainment-referencing
Construct assessed	Manifest performance	Latent ability	Underlying attainment
Standard setting	Criteria articulated in advance	Cut-scores established in advance, on the basis of a norming study	Grade boundaries established after the assessment is administered using a combination of statistical evidence and examiner judgement
Assessment methods	Open (authentic assessment of the set criteria is important)	Constrained (secure assessment under controlled conditions)	Range of possible assessment methods
Transparency of assessment	Comparatively *high*—all criteria known in advance	Comparatively *low*—test is kept highly secure and not released to learners or teachers	*Medium*—current examination papers kept secure until administration but past papers available
Potential negative washback effects on pedagogy	Pedagogy can be supplanted by setting of tasks and feedback regarding the pre-defined criteria	Pedagogy can involve knowledge of the likely assessment tasks and cramming to help the learner compete with others	Pedagogy is curriculum-focussed with assessment performance in mind, which can involve 'teaching to the test'
Score meaning	Performances meet the pre-specified criteria	Performance with respect to a reference population	Performance in relation to the curriculum, given the context in which assessment occurs (e.g., the assessment difficulty that year)
Paradigm	Outcomes-based	Psychometrics-based	Curriculum-based
What do I have to do to get the grade?	Meet the criteria	Achieve a specific score in the test, performing better than a specified percentage of the population norm	Demonstrate required knowledge and skills, with compensation for the difficulty of the task and context in which assessment is conducted

have encompassed the oversight and assurance of qualification standards, providing us with practical insights into the challenges associated with maintaining validity across diverse qualification types. We have been involved in many of the debates regarding the limitations of the current systems and the pros and cons of change. We have also influenced and overseen shifts in the standard setting processes over many decades of experience. This positionality is crucial in the context of the questions this article seeks to address, and the approach taken to analysis and interpretation of the literature. Our positions mean that we have an interest in the transparency of assessment at different levels of national systems—for the learner, teacher, school and institutions that use assessment results (Baird et al., 2018, 2024). In this article, we focus on transparency for teachers and learners. The authors are to some degree assessment industry/policy insider researchers (Grey, 2020) and hold academic positions in which they research and critically appraise standard setting systems internationally. This dual perspective strengthens the analysis of standard setting transparency by combining practitioner knowledge with academic rigour, enhancing both contextual sensitivity and the practical relevance of the findings.

While this professional proximity allowed for an informed and contextually grounded interpretation of the literature, it also required ongoing reflexivity to avoid uncritically reproducing the assumptions and values embedded within the assessment system. In engaging with the literature, we were particularly mindful of the tendency within the assessment community to prioritise reliability and stability of grading outcomes over other aspects of quality, such as construct validity and ease of interpretation by users of grades. We sought to balance this by engaging with critical perspectives, including from stakeholders, that challenge the assumptions of technical measurement.

Conceptualisation of qualification standards

Within research literature, the emphasis in discussions of standards is usually upon performance standards—those generated in the standard setting process when grade boundaries are set—after assessments have been devised, administered and marked (Opposs & Gorgen, 2018). This is an insider, technical perspective. Teachers and learners see qualification standards more broadly and tend to focus on content standards, which specify knowledge and skills to be learned. In evaluating approaches to standard setting, we take a holistic approach—considering standards as embedded in the end-to-end design and delivery of qualifications. Further, our focus is on the role of transparency in standard setting for raising educational standards rather than in relation to the multitude of other possible purposes to which assessment outcomes may be put (Newton, 2007a).

SCORE MEANING AND STANDARD SETTING APPROACHES

How scores can be interpreted is affected by the standard setting approach that is used and the philosophy behind it. Accordingly, we outline the current approach to standard setting for GCSEs and A-levels—attainment-referencing—and then explore whether norm- or criterion-referencing could provide greater clarity to teachers and learners regarding the expected standard. The likely pros and cons of any change of approach for GCSEs and A-levels are outlined.

Attainment-referencing

The definition of standard setting in use for GCSEs and A-levels is 'attainment-referencing' (Newton, 2022). It is a holistic approach which considers a range of information, including statistics, examiners' judgements of the difficulty of the assessments, examiners' judgements of the quality of learners' performances, and relevant research and policies (e.g., there may be a requirement to align standards between qualifications). The aim is to identify the underlying attainment of learners through consideration of these multiple sources of evidence. Taylor and Opposs (2018) provided a detailed account of how these multiple sources of evidence are brought together to set grade boundaries in A-level qualifications.

Attainment-referencing permits adaptation to events, such as a question paper that turns out to be too easy or a change in assessment policy. For example, adjustments to grading standards may be made in light of public concern about a subject's difficulty (Ofqual, 2018). The term was first used by Newton (2011), who noted that the approach had been previously described as 'weak-criterion referencing' (Baird et al., 2000) because it relies on examiner judgement (moderated by statistics) to consider the general quality of learner performance required for each grade in the context of the difficulty of the assessment.

Whilst the flexibility of attainment-referencing has benefits, it can be difficult for teachers and learners to understand exactly what the standard is and the reasons for any changes to that standard. For example, the attainment-referenced standard setting process accounts for the 'sawtooth effect', whereby cohort performance drops after a qualification has been reformed and then steadily improves over time, with lower grade boundaries set in the early years of delivery of new GCSEs and A-levels (Cuff et al., 2019; Linn et al., 1990). This information would almost certainly be of interest and use to teachers. However, it is not easy to explain and so it is not well understood. It contributes to the belief that the standard is ever-shifting—less than half of respondents to Ofqual's annual survey of perceptions of GCSEs and A-levels thought that standards were maintained year on year (Ofqual, 2024).

Norm-referencing

It is helpful to begin by describing the typical norm-referencing approach. Isaacs et al. (2013) suggested that:

> An educational assessment procedure can be identified as norm-referenced when the score that an individual achieves is converted into a statement or grade indicating how that individual compares with others who have undergone the same assessment.
>
> (p. 97)

Essentially, learners receive a score and/or grade that represents their rank in relation to the population of learners for whom the assessment is designed, including those who may have taken the assessment at a different time. To achieve this, the assessment is calibrated through research studies, called 'norming studies', undertaken with representative samples of the population of learners. Once a population-level distribution of scores is established, the scores of new test-takers may be mapped against it to ascertain their relative attainment (Shepard, 1979). Scores can then be translated into grades, if desirable.

In the UK, norm-referencing often enters public discourse (Geisinger, 2021), and many people mistakenly believe that qualifications are norm-referenced (Newton, 2022). Despite this, and the relative familiarity of the term, we were unable to find any international examples of it being used as the main approach to setting and maintaining standards for high-stakes national qualifications. Norm-referencing approaches are widely used in high-stakes testing contexts around the world, but rather than assessing the full breadth and depth of a curriculum, they use a psychometric approach to focus on a sharply defined construct. They often assess skills that correlate with success in higher education or in a specific job role; for example, college admissions tests in the United States (ACT, 2023; The College Board, 2017). Despite there being a lack of strongly comparable examples to draw upon, discussing the implications of norm-referencing in terms of score meaning for GCSE and A-level is helpful from a theoretical point of view. Would a shift to this approach to setting and maintaining standards make the link between learning and achievement more transparent for teachers and learners?

To successfully maintain a norm-referenced standard, a large bank of pre-tested items with known psychometric properties (e.g., item difficulty and discrimination) is required (Wright & Bell, 1984). Such item banks need to be confidential and secure, as the test would no longer provide a fair comparison to the norm if the population were to become familiar with the items. Familiarity would threaten the validity of the test and may also undermine its relationship with good teaching and learning by incentivising rote learning. Of course, this level of test security might also obfuscate score meaning for teachers and learners, making

it difficult for them to know exactly what is required of them for a particular grade. As we have outlined, there is a distinction to be made between the format, style and broad content of a test being familiar to learners and the questions themselves being predictable (Holmes et al., 2020).

Norm-referencing may also limit the availability of different methods of assessment. Coursework is generally incompatible with norm-referencing because it becomes increasingly familiar over time (Opposs, 2016). Again, this familiarity could allow teachers to support learners in a formulaic way, causing them to achieve higher scores in relation to the established norm in a manner that does not reflect improved attainment but rather improved assessment technique. This suggests that only constructs amenable to examination can be norm-referenced.

Norm-referencing sits most comfortably in the psychometric paradigm (Baird & Opposs, 2018). This approach emphasises scoring learners against a single scale intended to represent the core construct of interest. In GCSEs and A-levels, the construct would relate to 'ability' in the relevant subject. Psychometric approaches usually impose a normal distribution on two parameters: the item (how difficult it is) and the person (their ability with regard to the construct). The most familiar example of norm-referencing is in the context of intelligence testing. Here, an individual's score on a specific test is compared against an established distribution of scores for the population (Geisinger, 2021; Norfolk et al., 2015). The scores on intelligence tests are usually standardised such that the average score is 100 and the standard deviation is 15. Due to the statistical properties of the normal distribution, approximately two-thirds of the population will score between 85 and 115. Similarly, the aforementioned norm-referenced college admissions tests often convert scores into percentile ranks, stating that, for example, a score of 30 out of 36 places the learner in the 95th percentile for Maths (ACT, 2023).

Perhaps because of this, norm-referencing is often confused with cohort-referencing (Newton, 2022; Wiliam, 1996), whereby a predefined proportion of learners is awarded each available grade following each assessment period. For example, the highest attaining 10% of learners receive an A grade, the next highest attaining 10% receive a B grade, etc. Unlike in cohort-referencing, norm-referencing allows for the proportion of learners getting each grade each year to change. If the cohort in 1 year performed particularly well, this would be reflected in their scores relative to the established norm, and they would achieve a higher average grade. Similarly, if a cohort were to underperform, perhaps because of a change in national policy or a significant national disruption, such as the closure of schools during the COVID-19 pandemic, that cohort's average grade would be lower than that of past cohorts.

Under norm-referencing, grade boundaries remain static, a product of the normal distribution that has been established in the norming study population and a decision at that stage about what proportion of that population deserves each grade. A core principle at the heart of norm-referencing is that:

> …the raw scores are a manifest expression of a latent personality trait or ability which itself cannot be directly assessed… norming aims at mapping the raw scores of a test to that latent ability. While the latter one is usually assumed to be normally distributed, the same unfortunately does not apply to the raw score distribution.
>
> Lenhard et al. (2019, p. 2)

This principle, that the test is evaluating a learner on a (latent) construct which we expect to be normally distributed throughout the population, is crucial to interpreting the meaning that grades have in such a system. A grade would not be reported as an individual's attainment in relation to the curriculum; rather, it would represent their attainment relative to the

population. For example, a grade C might represent attainment within the 40th and 60th percentile of the population. Whether learners had attained a particular performance standard in, for example, algebra or creative writing would not be explicitly referenced.

Under a norm-referenced approach, how would one respond to a learner asking the question, 'what do I need to do to get a Grade A?' The candid answer might be: 'you need to get a better score than 80% of the norming population who took this exam'. A follow-up question might then be 'how do I do well on the test?' One possible answer may lie in an analysis of test data to identify the items which typically discriminate between candidates at different grades (Greatorex et al., 2001). This might allow test developers to provide a profile that describes typical performance at each grade.

Though potentially helpful, this approach would encounter an issue that exists with any scored examination—there are many different routes through which a learner may achieve their overall mark. GCSEs and A-levels are compensatory assessments (Wiliam, 1995). Two learners who achieve the same mark may have performed quite differently across the subject content, meaning that their knowledge and skills will not be identical. A learner who achieves 50% of the available marks may do so by gaining half of the available marks for every item or by performing perfectly on half of the items while receiving zero marks for the other half. For this reason, any descriptions of 'typical' performance would be necessarily vague or probabilistic (as the descriptor will apply to most, but not all, performances), as is the case for existing GCSEs and A-level grade descriptors (Cadwallader, 2014). This is because, as with attainment-referencing, the assessment standard in a norm-referenced system does not, in a fundamental sense, directly reference the curriculum. In contrast, the criterion-referencing approach to setting standards promises to provide a direct link to the curriculum (Millman, 1994). We consider this approach next.

Criterion-referencing

In terms of Baird and Opposs' (2018) assessment paradigms, criterion-referencing aligns most closely with an outcomes-based paradigm. This approach originated in occupational training and was linked to the development of outcomes-based learning (Wolf, 1995) in which goals, outcomes, or criteria describing what learners must know and be able to do are established in advance of any assessment.

The term 'criterion-referenced' was introduced by Glaser (1963) to frame a debate about test design and the interpretation of scores. Criterion-referenced assessments were intended to increase the fidelity of score meaning:

> Measures which assess student achievement in terms of a criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others.
>
> (Glaser, 1963, pp. 159–160)

In a pure criterion-referenced system, sometimes referred to as 'strong' criterion-referencing (Croll et al., 1994; Stringer, 2012), it is in the setting of standards that criterion-referencing differs most from attainment-referencing. This is because standards are not determined by a post-hoc standard setting process, but are defined in advance, often using a combination of high-level verbal descriptions of the skill to be demonstrated (variously referred to as outcomes, objectives, standards) and more detailed descriptions of the level of performance that would merit a particular grade (often referred to as performance or assessment criteria). These outcomes and performance criteria are defined by subject

experts, often through a committee or consultation process that attempts to determine the most important skills and knowledge to be learned in the subject.

Criterion-referencing is widely used in education systems around the world. For example, in the French Baccalaureate (Gauthier, 2018), the Swedish teacher assessments (Wikström & Pantzare, 2018), and the Queensland teacher assessments (Campbell, 2018). Sometimes the term itself is used, and sometimes words like 'competence', 'standards' or 'outcomes' describe the approach. Often its introduction is associated with moves to outcomes-based school curricula and a policy intention to modernise the curriculum, placing less emphasis on knowledge and more on skills and competences, for example, in Norway (Tveit, 2014) and Ontario (DeLuca et al., 2017; Ontario Ministry of Education, 2010).

The design of systems such as these, used in general education contexts, has features that are distinct from those of strong criterion-referencing. Their design aligns more closely to an approach first defined by Sadler in the 1980s—'standards-referenced assessment' (Sadler, 1987). It allows reporting of learners' actual achievements, whilst acknowledging the centrality of judgement in the assessment process, thus creating greater space for teacher input into assessment decisions. As variations of this approach have been implemented globally, some common features can be seen:

- Learner performance is judged against criteria, sometimes envisaged (and set out) as existing on an underlying continuum of quality.
- Criteria take the form of descriptions of acceptable performance, defined in advance, by subject experts.
- The criteria can be used as part of the learning process, to help learners to understand their progress.
- Assessment consists of a series of tasks or assessments, or at least on more than a one-off final exam; within or across those tasks, assessment decisions can comprise of holistic or best fit judgement of learner proficiency.
- This means that coursework tasks play a role in determining the final assessment result, sometimes assessed directly by the learner's teacher.

(adapted from Sadler, 1987)

Popham (2014), one of the foremost advocates of criterion-referenced assessment, foregrounded the intrinsic link between a criterion-referenced approach and the curriculum as a key strength, emphasising the potential benefits to teaching and learning. Indeed, criterion-referencing can be viewed by policymakers as a way to ensure that learning programmes cover skills and knowledge that they define as essential. For example, in Estonia, decisions on curriculum content are largely a matter for schools, but the Ministry of Education has published a detailed Digital Skills framework for learners. Whilst not mandatory, this framework sets out detailed outcomes and performance criteria to support schools in developing digital skills seen as necessary for the 21st century (Republic of Estonia Ministry of Education and Research, 2023).

Another purported key strength of a criterion-referenced approach is that it is designed to be transparent—for assessors, teachers, learners, parents and users of the qualification. Statements of standards are usually shared with learners, so that they know what they must do and have agency in their learning progress (Balloo et al., 2018). Similarly, for those who want to know what the assessment result means, they can read the statement of standards and know what skills and knowledge the learner has had to demonstrate to achieve the grade.

In theory, criterion-referencing appears to closely align the curriculum with the assessment standards. By definition, the standards that the learner has to achieve are specified up front, before learning or assessment begins, and in terms that are intended to be transparent.

In this way, criterion-referencing, in principle at least, answers both the learner's question about what they have to do to achieve the grade and the assessment user's question about what the grade means in terms of what the learner knows and can do.

However, in practise, criterion-referencing poses significant challenges that can undermine the relationship between the criteria, learning and score meaning. The primary challenge is avoiding over-prescription and a lack of curriculum flexibility (as outlined by Sizmur and Sainsbury (1997) in the context of the National Curriculum in England and Wales). While skills and standards are specified in advance, the content and context are intentionally left flexible—allowing teachers to tailor them to local circumstances or even personalise them for individual learners, as exemplified by the system in Norway (Nusche et al., 2011; Tveit, 2014).

Though this may be welcome to teachers from a curriculum point of view, when they are required to assess learner progress and are aware of the stakes, rather than experiencing curricular flexibility as freedom, teachers can experience it as vague and unhelpful (Sizmur & Sainsbury, 1997). This can lead to a cycle of ever-increasing detail being added to assessment requirements, often with new jargon invented along the way. For example, an OECD (2021) review commented on Scotland's Curriculum for Excellence (CfE):

> The evolving CfE seems overloaded with numerous elements: the vision around its four capacities (with attributes and capabilities); seven principles; eight curriculum areas; curriculum entitlements; qualifications; expectations and outcomes; benchmarks; moderations; progression levels; and more.
>
> (p. 57)

As critics of criterion-referenced assessment have argued (Cresswell & Houston, 1991; Wolf, 1995), the attempt to define requirements more precisely can never reach a point of sufficiency. The standards—of both content and performance—will always be opaque, or even inconsistent, not because assessment developers have done a poor job, but because their task is impossible. Cresswell (1987) gives the example of a National Curriculum outcome in England, in which the defined competence is that the learner must 'use conventional punctuation':

> How is the phrase "use conventional punctuation" to be interpreted operationally? How many errors of punctuation can be permitted before "use" has not been established? Can a candidate who never uses inverted commas but otherwise punctuates perfectly be deemed to have met the criterion? And so on.
>
> (Cresswell, 1987, p. 250)

These problems are exacerbated when the assessment has to provide a level or grade for the learner's achievement. Discussing language assessments in South Africa, Allais (2012) argued that the same competence statements could be perceived as applying to widely differing levels of performance. For instance, the outcome 'demonstrate awareness of manipulative devices' may be evident in primary school pupils (e.g., through nursery rhymes), in newly literate adults (e.g., through understanding simple slogans), and in individuals using language at an advanced academic level.

Interpreting the language of criteria can be difficult, if not impossible. Knight (2001) argued that certain forms of learning—particularly those of a complex nature—resist reduction to precise formulations that clearly predict outcomes. He used the following example to explicate the problem: to assert that a desired outcome of undergraduate education is the capacity to act with a high degree of autonomy raises immediate questions about what is meant by both *autonomy* and *a high degree*. Even detailed inventories of intended outcomes

fail to delineate the scope of such attainment. Is this autonomy, however defined, to be demonstrated exclusively within academic study, or is it also assumed to extend to professional contexts and to life more generally, across circumstances that may be simple or complex, familiar or unfamiliar? Moreover, what criteria determine whether a situation is to be regarded as *familiar*, *simple* or otherwise?

Language issues such as these make it difficult for those setting content standards to do so consistently and effectively. The same issues of interpretation mean that the criteria can be problematic to use: performance standards themselves are not transparent and the resultant assessment decisions are inconsistent (Wikström & Pantzare, 2018). This poses a fundamental threat to the validity of the assessment (Bloxham et al., 2011). Critics have argued that no amount of detail can remove the need for assessor judgement, made in the particular context, not just of the curriculum content, but also of the particular assessment task on which the learner is being judged. Instead, some have advocated for a more interpretive or sociocultural approach to understanding standards, recognising that criteria acquire meaning only when educators interpret them through their own personal frameworks and perspectives (Bloxham et al., 2016). Ajjawi et al. (2021), on the other hand, advocated for a sociomaterial perspective on assessment standards. They argued that standards are not fixed but are dynamically co-produced through interactions between people and material artefacts, such as rubrics, within specific contexts. Whichever perspective is taken, it is clear that the interpretation of assessment criteria is far from unproblematic.

Another issue with criterion-referenced approaches to standards is the potential for volatility in cohort outcomes. As with a norm-referenced approach, a criterion-referenced approach does not seek to ensure any degree of consistency in terms of the proportions of learners who achieve each grade each year. Passing a given grade is a matter of meeting the required assessment criteria—if every learner fails to meet the criteria, then none of them will receive the grade. If every learner achieves the criteria, the whole cohort will receive the grade. This potential volatility in outcomes could have implications for those who use the qualifications to make decisions about employability or access to higher education. For example, in New Zealand, fluctuations in the proportion of learners achieving grades in standards-based qualifications have proved controversial (Isaacs & Gorgen, 2018).

Aggregation and differentiation can also cause tensions with the use of criterion-referencing in national qualifications. Often employing conjunctive rather than compensatory aggregation (Wiliam, 1995), the approach tends to be associated with the assessment of whether or not competence has been achieved, rather than the differentiation of performance across a wider range of grades (grades A* to E, for example). Providing a wider range of outcomes than 'Pass' and 'Fail' is possible but requires finer judgements from assessors and substantively more detailed assessment criteria. The latter risks knowledge and skills becoming overly specified and atomised, with key areas of the curriculum being specified ever more precisely to allow for differentiation between grades.

FURTHER ISSUES IN APPLYING CRITERION- AND NORM-REFERENCING IN GCSES AND A-LEVELS

Both criterion- and norm-referencing approaches to standard setting are designed to enable certain conclusions about a learner's performance to be drawn, whether it be in relation to specified criteria or to an established norm. As such, under either system, grades apparently signal a strong meaning, something that stakeholders desire. Why then do GCSEs and A-levels take an attainment-referenced approach?

It is telling that while examples of norm-referenced tests exist, we have been unable to identify a national school-leaving qualification that takes a norm-referencing approach to standard setting. The requirement that tests are kept secure after they are taken would be likely to cause significant friction within an assessment culture in which the desire for transparency is high, and teachers and learners use previous papers in exam preparation. The regular redesign of the curriculum and associated reform of qualifications (Baird & Lee-Kelley, 2009) would require frequent norming studies. Indeed, a norm-referencing approach does not fit easily with the flexibility needed to maintain an up-to-date curriculum. The need for teacher assessment of those elements of the curriculum that cannot be validly assessed by exam is another problem (Opposs, 2016).

Norm-referencing fits best with a psychometric paradigm that views exam marks as the manifestation of latent ability that is expected to be normally distributed throughout the population (Lenhard et al., 2019). This is somewhat at odds with the current operation of qualifications within a curriculum-embedded paradigm that sees learning as malleable, even if distributed in the population (Baird & Opposs, 2018).

Further, norm-referencing is often misunderstood by teachers and the public and is sometimes misunderstood or defined imprecisely in the literature. It is regularly confused with cohort-referencing (Newton, 2022). It is debatable whether norm-referencing would be publicly supported were it properly understood and whether the notion of comparison to the performance of a past group of learners (the 'norm') would be acceptable. While GCSE and A-level stakeholders are keenly interested in historical comparison and the maintenance of standards over time (Newton, 2007b), there is also a desire to be flexible and consider current context. For example, the adjustment to the standards expected during the early awards of reformed qualifications, so that learners were not unfairly penalised due to teachers' limited familiarity with the new course content and assessment style, was welcomed (Cuff et al., 2019). Further, to account for the uneven impact of school closures during the COVID-19 pandemic on learning, the approach taken to standard setting in 2021 was significantly modified. Learners were assessed only on the content they, individually, were taught (see Ofqual, 2022).

When GCSEs were introduced in 1988, there was an aspiration for them to be criterion-referenced (DES, 1987). However, an unmanageable volume of criteria were produced for many subjects, which fuelled the fear that criterion-referencing would negatively affect teaching and learning through over prescription (Sizmur & Sainsbury, 1997). Research also showed that despite efforts to clearly specify the criteria, they were interpreted differently by examiners (SEC, 1984).

Moreover, there are other features of strong criterion-referencing which sit uneasily with GCSEs and A-levels. The assessment encompasses the entire curriculum, and learners are required to meet the standard across all components under strong criterion-referencing; high performance in one area cannot offset underperformance in another. In effect learners are graded on their weakest performance—the failure to meet a single criterion will result in a grade not being awarded (Forrest & Shoesmith, 1985). Perversely, this could impact the reliability of grading of learners—for example, a very able learner might slip up on a single, trivial question, which might result in them being awarded a lower grade than if some level of compensation across the exam were permitted.

If the maintenance of the proportion of learners achieving each grade each year is desirable, any move to a strongly criterion-referenced system would present a risk to the stability of the system. Without careful calibration of the criteria, such that none were overly difficult or overly easy to achieve, there would be potential for catastrophic unintended effects on overall outcomes. For example, the proportion of learners reaching the highest grades may decline dramatically if just one criterion was too difficult to achieve. Equally, should criteria be too easy to achieve, outcomes could rise such that very few learners would receive lower

grades and differentiation would become problematic. The public are likely to question the reliability or legitimacy of the system if national outcomes fluctuate substantially in ways that defy straightforward explanation or are perceived to be unfair.

HOW APPROACHES TO STANDARDS INTERACT WITH LEARNING, TEACHING AND THE CURRICULUM

Transparency in score meaning differs across the standard setting approaches discussed. In criterion-referenced approaches, the aim is to signal to learners exactly what knowledge, skills and behaviours are required. Generalising is less of an aim. Being able to land a plane, conduct a heart operation or wire a kitchen are sufficient criteria in themselves and there is usually little expectation in an outcomes-based approach that scores need to generalise to different contexts and situations, in theory. Of course, in practise, educational assessment is not solely about replicating the same performances under the same conditions. Additionally, in school and college settings, most courses anticipate a broader curriculum than that encapsulated in a list of criteria and the capacity to generalise from scores to performances in other contexts.

In terms of teaching and learning, criterion-referencing aims for such a clear exposition of the criteria that pedagogy can be supplanted. Torrance's (2007) research across a range of vocational qualifications found that criteria compliance had replaced the normal conception of teaching and learning, leading him to term this 'assessment *as* learning'. Learners would be familiarised with the criteria and sent off to produce work that demonstrated their mastery. Lecturers would then check off the lists to accredit their learning. The experience for learners was demotivating and the value-added by the educational institution questioned. Learning was reduced to an instrumental process of production. This kind of reductionist approach, which conflates the curriculum, learning and teaching to the assessment criteria, is a reality in some settings.

Dann (2012), however, conceptualised assessment *as* learning very differently—as an approach that integrates assessment seamlessly into the learning process, emphasising the active role of learners in their own educational journey. There is a risk for GCSEs and A-levels that the combination of school-accountability pressures and increased transparency would lead to highly instrumental approach to learning rather than the positive outcomes envisaged by Dann. Whilst score meaning is palpable under criterion-referencing, this level of transparency may not be desirable for broad, general education (Torrance, 2007). Indeed, too much specification may not be helpful for learners' development of higher order thinking skills in higher education either.

Norm-referencing has a different relationship with transparency. The tests themselves are held secure, precisely so that people cannot teach to the test. To do so would disrupt the statistical properties of the items and overall test scores. The items would be easier for the test-takers who had been prepared directly for them. Thus, the norms would no longer be valid. A test-taker at the 90th percentile who had prior knowledge of the content of the test and had been specifically prepared for it would likely have scored at lower than the 90th percentile under less transparent conditions. Therefore, the 90th percentile score tells us something different about the well-prepared test-taker than it would about those scoring at the 90th percentile in the norming study.

In unpicking this relationship, the assumptions about score meaning and its relationship with learning and teaching become apparent. In a norm-referencing tradition, the learning and teaching are separate from the test, which merely assesses the extent of learning. Fundamentally, the learner has a particular ability level that the test uncovers. Whilst preparation is intended, the idea is that it should be a general preparation rather than specific to

the items in the test, so that the results can be generalised to likely performances elsewhere. Norm- and criterion-referencing therefore have very different relationships with transparency, teaching, learning and the curriculum.

That said, most national qualifications have attempted to embrace the benefits of using statistical information in standard setting (to improve stability) and production of criteria or progress scales (to provide clarity). For this reason, the prototypical example of a norm-referenced achievement test, the US SAT, incorporated criteria in the 1970s and eminent researchers claimed that it was reasonable to call the test both criterion- and norm-referenced (Angoff, 1974). Equally there are recent attempts by eminent intelligence researchers to produce criterion-referenced assessments of intelligence (Sternberg et al., 2022). Attainment referencing, as applied to GCSEs and A-levels, is such a blended approach and faces the problem that some users interpret the score as representing a learner ability, whilst others interpret it as a performance in relation to criteria.

Each of the three approaches to setting and maintaining standards discussed in this article have different implications for score meaning. As we have seen, in both theory and practise, no approach is able to deliver every expectation that may be placed upon it, hence the importance of policy makers and stakeholders understanding the nuances and compromises inherent to each approach. Table 1 summarises some of the key points that have been discussed in this paper for criterion-referencing, norm-referencing and attainment-referencing. Note that higher education typically follows a criterion-referenced, outcomes-based approach to assessment. There are exceptions to this, when psychometric approaches are used, for example, in admissions testing, some medical examinations and so on. However, the promise of transparency through criterion-referencing that falls short for school qualifications also fails to deliver for higher education, for the same reasons (Bearman & Ajjawi, 2018).

WHAT CAN BE DONE TO IMPROVE SCORE MEANING IN ATTAINMENT-REFERENCED QUALIFICATIONS?

Gonsalves and Lin (2024), writing about assessment practises in UK higher education, argued that improvements in transparency require a holistic strategy, incorporating techno-rational, sociocultural and sociomaterial approaches. Transparency necessitates explicit documentation, for example describing the assessment criteria, but it also requires opportunities for dialogue, collaboration and shared understanding among stakeholders, along with supported opportunities for stakeholders to engage with material artefacts to shape their understanding.

Fortunately, there are different ways to bolster teacher and learner understanding of what is required to achieve a given grade. First, a fruitful avenue may be to empirically identify examination items that typically discriminate between candidates at different grades (Greatorex et al., 2001), allowing awarding bodies to provide post-hoc grade descriptors describing typical attainment at each grade. Given that examination items are different for each cohort, there would be limitations to the usefulness of such information, but if such caveats were explained there would still be value for teachers and learners. Second, carefully anonymised examples of candidate assessment evidence from across the grade range could be made available. Such materials would complement grade descriptors, not least by demonstrating some of the many different routes to a grade. However, as Gonsalves and Lin (2024) argued, it would not be enough to merely create and disseminate these kinds of materials to teachers. More social opportunities would need be created across schools and colleges for teachers to interact with these artefacts, to develop a consensus on standards, and to clarify the performances that satisfactorily meet those standards. Furthermore, teachers

would need to be supported to engage their learners in the process of understanding the assessment criteria and, broadly speaking, the standards required to achieve grades.

CONCLUSION

No approach to standard setting would provide teachers and learners with absolute clarity when it comes to score meaning. It is evident from the literature that this is also true in higher education. Complete transparency of expected standards in either a school or higher education context is an unrealistic aspiration. It is nonetheless important to continue to strive to better signal the complex and tacit performance expectations, while recognising the trade-offs that different approaches may bring. In the context of GCSEs and A-levels, alternative approaches to attainment-referencing may be more transparent in providing score meaning, but they may also have significant implications for assessment validity and, crucially, for supporting teaching and learning. The details of any standard setting process are honed over many years of experience. Approaches are culturally embedded, and major paradigm shifts are infrequent and high risk (Isaacs, 2018). There are reasons why qualifications such as GCSEs and A-levels are not norm- or criterion-referenced and why an attainment-referenced approach is used.

The strength of attainment-referencing is that the performance required to achieve a grade depends upon the difficulty of the task set. This is important to the fairness of GCSEs and A-levels. But it is a drawback of the approach too, as the performance required varies over time. This, combined with the compensatory nature of the examinations, weakens the conclusions that can be drawn about what has been achieved for a particular grade. This lack of clarity can translate to less certain teaching and learning. However, while the current system of attainment referencing may mean that learners cannot *precisely* know what it is they need to do to achieve a grade, there are opportunities to improve the channels and methods of communication to increase clarity. Better use of the available data, specifically to explore whether post-hoc grade descriptors would be valuable, and greater opportunity for social engagement whereby teachers are able to understand and discuss procedures for setting and maintaining standards, are two such opportunities. The imminent reform of GCSEs and A-levels provides opportunity for the integration of such approaches from the outset.

METHODOLOGICAL LIMITATIONS

As a critical integrative synthesis undertaken by professionals with experience of the policy and practise of standard setting, this study inevitably involved interpretive subjectivity. The process of selecting, analysing and integrating literature on standard setting was shaped by the authors' insider understandings of how GCSEs and A levels operate in practise, and by their professional interest in promoting transparency for learners and teachers. While such expertise provides depth, contextual sensitivity and a grounded appreciation of operational realities, it also risks reinforcing established perspectives within the system. This interpretive subjectivity is intrinsic to the kind of inquiry undertaken here, which seeks to connect theory and practise rather than to claim neutrality. To manage it, efforts were made to maintain reflexive awareness of our own assumptions and prejudices, and to explicitly justify our inferences about the pros and cons of different approaches. The resulting synthesis is a theoretically informed and professionally situated interpretation of how the literature might inform new approaches to standard setting in the UK and for other national assessment systems using or considering these definitions of standards.

AUTHOR CONTRIBUTIONS

Stuart Cadwallader: Writing – original draft; conceptualization; writing – review and editing. **Lena Gray:** Writing – original draft; conceptualization; writing – review and editing. **Jo-Anne Baird:** Writing – original draft; conceptualization; writing – review and editing. **Michelle Meadows:** Writing – original draft; conceptualization; methodology; writing – review and editing.

FUNDING INFORMATION

This research was supported by Qualifications Wales, UK (grant number QW222313).

CONFLICT OF INTEREST STATEMENT

None of the authors believe there to be any conflicts of interest arising from this research.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

ETHICS STATEMENT

The research outlined in this manuscript is a critical evaluation paper based entirely on published literature. It did not involve the collection of data from human participants, and we therefore did not undertake a formal ethics approval process. We have ensured that our research complies with BERAs ethics guidelines for education research throughout the process.

REFERENCES

ACT. (2023). *ACT test scores: National ranks*. ACT. https://www.act.org/content/act/en/products-and-services/the-act/scores/national-ranks.html

Ajjawi, R., Bearman, M., & Boud, D. (2021). Performing standards: A critical perspective on the contemporary use of standards in assessment. *Teaching in Higher Education*, *26*(5), 728–741. https://doi.org/10.1080/13562517.2019.1678579

Allais, S. (2012). Claims vs. practicalities: Lessons about using learning outcomes. *Journal of Education and Work*, *25*(3), 331–354. https://doi.org/10.1080/13639080.2012.687570

Alliance for Workers' Liberty. (2021). *Abolish GCSEs, turn the tide on toxic testing*. Workers' Liberty – Reason in Revolt. https://www.workersliberty.org/blogs/lewisham-teacher/2021-11-23/abolish-gcses-turn-tide-toxic-testing

Angoff, W. H. (1974). *Criterion-referencing, norm-referencing, and the SAT (RM-74-01; ETS Research Memorandum)*. https://www.ets.org/research/policy_research_reports/publications/report/1974/imcl.html

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*(5), 258–267. https://doi.org/10.3102/0013189X07306523

Baird, J., Godfrey-Faussett, T., Allan, S., MacIntosh, E., Hutchinson, C., & Wiseman-Orr, L. (2024). Standards as a social contract in curriculum-based qualifications: Stakeholder views in Scotland. *Cambridge Journal of Education*, *54*, 455–474. https://doi.org/10.1080/0305764X.2024.2377965

Baird, J.-A., Ahmed, A., Hopfenbeck, T., Brown, C., & Elliott, V. (2013). *Research evidence relating to proposals for reform of the GCSE*. https://ora.ox.ac.uk/objects/uuid:7816d592-0dd6-4882-acf7-72fbd16decd2/files/m3f7d3a9d81a5eb426eb9b8c46844d37b

Baird, J.-A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: Fields apart? *Assessment in Education*, *24*, 317–350. https://doi.org/10.1080/0969594X.2017.1319337

Baird, J.-A., Cresswell, M., & Newton, P. (2000). Would the real gold standard please step forward? *Research Papers in Education*, *15*(2), 213–229. https://doi.org/10.1080/026715200402506

Baird, J. A., & Lee-Kelley, L. (2009). The dearth of managerialism in implementation of national examinations policy. *Journal of Education Policy*, *24*(1), 55–81. https://doi.org/10.1080/02680930802382938

Baird, J.-A., Isaacs, T., Opposs, D., & Grey, L. (Eds.). (2018). *Examination standards: How measures and meanings differ around the world*. UCL, IOE Press.

Baird, J.-A., & Opposs, D. (2018). Chapter 1: The standard setting project: Assessment paradigms. In J.-A. Baird, T. Isaacs, D. Opposs, & L. Grey (Eds.), *Examination standards: How measures and meanings differ around the world*. UCL, IOE Press.

Balloo, K., Evans, C., Hughes, A., Zhu, X., & Winstone, N. (2018). Transparency isn't spoon-feeding: How a transformative approach to the use of explicit assessment criteria can support student self-regulation. *Frontiers in Education*, *3*(69), 1–11. https://doi.org/10.3389/feduc.2018.00069

Bearman, M., & Ajjawi, R. (2018). From "seeing through" to "seeing with": Assessment criteria and the myths of transparency. *Frontiers in Education*, *3*(96), 1–8. https://doi.org/10.3389/feduc.2018.00096

Benton, T. (2016). *Comparable outcomes: Scourge or scapegoat?* Cambridge assessment research report. Cambridge Assessment.

Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: The role of assessment criteria in UK higher education grading practices. *Studies in Higher Education*, *36*(6), 655–670. https://doi.org/10.1080/03075071003777716

Bloxham, S., Den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: Exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education*, *41*(3), 466–481. https://doi.org/10.1080/02602938.2015.1024607

Cadwallader, S. (2014). *Developing grade descriptions for the new GCSEs: Considerations and challenges (CERP Report)*. Centre for Education Research and Practice. https://filestore.aqa.org.uk/content/research/CERP_RP_SMC_14052014_0.pdf?download=1

Campbell, M. (2018). Chapter 10: Standard setting in Queensland: The Queensland certificate of education. In J.-A. Baird, T. Isaacs, D. Opposs, & L. Grey (Eds.), *Examination standards: How measures and meanings differ around the world*. UCL, IOE Press.

Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 212–225. https://doi.org/10.1080/0969594X.2015.1063479

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use* (1st ed.). Routledge. https://doi.org/10.4324/9780429291661

Cresswell, M. J. (1987). Describing examination performance: Grade criteria in public examinations. *Educational Studies*, *13*(3), 247–265. https://doi.org/10.1080/0305569870130305

Cresswell, M. J., & Houston, J. G. (1991). Assessment of the national curriculum—some fundamental considerations. *Educational Review*, *43*(1), 63–78. https://doi.org/10.1080/0013191910430106

Croll, P., Abbott, D., Broadfoot, P., Osborn, M., & Pollard, A. (1994). Teachers and education policy: Roles and models. *British Journal of Educational Studies*, *42*(4), 333–347. https://doi.org/10.2307/3121675

Cuff, B. M. P., Meadows, M., & Black, B. (2019). An investigation into the Sawtooth effect in secondary school assessments in England. *Assessment in Education: Principles, Policy & Practice*, *26*(3), 321–339. https://doi.org/10.1080/0969594X.2018.1513907

Daly, A. L., Baird, J. A., Chamberlain, S., & Meadows, M. (2012). Assessment reform: Students' and teachers' responses to the introduction of stretch and challenge at A-level. *Curriculum Journal*, *23*(2), 139–155. https://doi.org/10.1080/09585176.2012.678683

Dann, R. (2012). *Promoting assessment as learning: Improving the learning process*. Routledge.

DeLuca, C., Braund, H., Valiquette, A., & Cheng, L. (2017). Grading policies and practices in Canada: A landscape study. *Canadian Journal of Educational Administration and Policy*, *184*, 4–22.

Department for Education. (2025). *Curriculum and assessment review: Interim report*. https://www.gov.uk/government/publications/curriculum-and-assessment-review-interim-report

DES. (1987). *Improving the basis for awarding GCSE grades*. 1st Annual Conference of the Joint Council for the GCSE.

Forrest, G. M., & Shoesmith, D. J. (1985). *Monitoring standards in the general certificate of secondary education*. Joint Matriculation Board.

Gauthier, R.-F. (2018). Chapter 8: Standard setting in France: The baccalauréat. In J.-A. Baird, T. Isaacs, D. Opposs, & L. Grey (Eds.), *Examination standards: How measures and meanings differ around the world*. UCL, IOE Press.

Geisinger, K. F. (2021). The history of norm and criterion referenced testing. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 42–64). Routledge. https://doi.org/10.4324/9780367815318

Gipps, C. (1999). Sociocultural aspects of assessment. *Review of Research in Education*, *24*, 357–392.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, *18*, 519–521. https://doi.org/10.1037/h0049294

Gonsalves, C., & Lin, Z. (2024). Clear in advance to whom? Exploring 'transparency' of assessment practices in UK higher education institution assessment policy. *Studies in Higher Education*, *50*, 1–17. https://doi.org/10.1080/03075079.2024.2381124

Greatorex, J. (2002). Making accounting examiners' tacit knowledge more explicit: Developing grade descriptors for an accounting A-level. *Research Papers in Education*, *17*(2), 211–226. https://doi.org/10.1080/02671520210122892

Greatorex, J. (2005). A review of research about writing and using grade descriptors in GCSEs and A levels. *Research Matters: A Cambridge Assessment Publication*, *1*, 8–11. https://www.cambridgeassessment.org.uk/Images/507975-a-review-of-research-about-writing-and-using-grade-descriptors-in-gcses-and-a-levels.pdf

Greatorex, J., Johnson, C., & Frame, K. (2001). Making the grade—developing grade descriptors for accounting using a discriminator model of performance. *Westminster Studies in Education*, *24*(2), 167–181. https://doi.org/10.1080/0140672010240207

Grey, L. (2020). Evidence-based policy-making and exam board insider researchers: Creating communicative spaces. *Assessment in Education: Principles, Policy & Practice*, *27*(2), 142–159. https://doi.org/10.1080/0969594X.2020.1749557

Hayward, L., Baird, J., Godfrey-Faussett, T., Rhandawa, A., Allan, S., MacIntosh, E., Hutchinson, C., Spencer, E., & Wiseman-Orr, L. (2023). National qualifications in Scotland: A lightning rod for public concern about equity in the pandemic. *European Journal of Education: Research, Development and Policy*, *58*(1), 83–97. Special Issue: Trust in Standardised Assessments. https://doi.org/10.1111/ejed.12543

Holmes, S., Khan, A., Zanini, N., & Black, B. (2020). *Predicting predictability*. Ofqual. https://www.gov.uk/government/publications/predicting-predictability

Hudson, J., Bloxham, S., den Outer, B., & Price, M. (2015). Conceptual acrobatics: Talking about assessment standards in the transparency era. *Studies in Higher Education*, *42*(7), 1309–1323. https://doi.org/10.1080/03075079.2015.1092130

Isaacs, T. (2018). Chapter 16: Setting standards in national examinations: What we have learnt. In J.-A. Baird, T. Isaacs, D. Opposs, & L. Grey (Eds.), *Examination standards: How measures and meanings differ around the world*. UCL, IOE Press.

Isaacs, T., & Gorgen, K. (2018). Chapter 15: Culture, context and controversy in setting national examination standards. In J.-A. Baird, T. Isaacs, D. Opposs, & L. Grey (Eds.), *Examination standards: How measures and meanings differ around the world*. UCL, IOE Press.

Isaacs, T., Zara, C., Herbert, G., Coombs, S. J., & Smith, C. (2013). *Key concepts in educational assessment*. SAGE.

JCQ. (2021). *GCSE grade descriptors to assist with determining grades*. https://www.jcq.org.uk/wp-content/uploads/2021/04/Summer-2021-Grade-Descriptors-GCSE.pdf

Knight, P. T. (2001). Complexity and Curriculum: A process approach to curriculum-making. *Teaching in Higher Education*, *6*(3), 369–381. https://doi.org/10.1080/13562510120061223

Leech, T. (2023). Assessment in England at a crossroads: Which way should we go? *Research Matters*, *35*, 80–93.

Lenhard, A., Lenhard, W., & Gary, S. (2019). Continuous norming of psychometric tests: A simulation study of parametric and semi-parametric approaches. *PLoS One*, *14*(9), e0222279.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of the claims that "Everyone is above average". *Educational Measurement: Issues and Practice*, *9*(3), 5–14.

Lucas, B., Hyman, P., & McConville, A. I. (2020). 3 reasons GCSEs need to change – and 3 alternatives. *TES Magazine*. https://www.tes.com/magazine/archive/3-reasons-gcses-need-change-and-3-alternatives

McArthur, J. (2020). Learning from exam results crisis: The way students' work is assessed needs to change. *The Conversation*. https://theconversation.com/learning-from-exam-results-crisis-the-way-students-work-is-assessed-needs-to-change-144710

McManus, H. (2018). Chapter 9: Standard setting in Ireland: The leaving certificate. In J.-A. Baird, T. Isaacs, D. Opposs, & L. Grey (Eds.), *Examination standards: How measures and meanings differ around the world*. UCL, IOE Press.

Meadows, M., Baird, J., Grey, L., Cadwallader, S., Godfrey-Faussett, T., Saville, L., Debnam, C., & Stobart, G. (2023). *Standards in GCSEs in Wales: Approaches to defining standards. OUCEA/23/1*. https://www.education.ox.ac.uk/research/research-on-standards-in-gcses-in-wales/

Millman, J. (1994). Criterion-referenced testing 30 years later: Promise broken, promise kept. *Educational Measurement: Issues and Practice*, *13*(4), 19–39.

Newton, P. (2007a). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 149–170. https://doi.org/10.1080/09695940701478321

Newton, P. (2007b). Contextualising the comparability of examination standards. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. QCA Press.

Newton, P. (2022). Demythologising A level exam standards. *Research Papers in Education*, *37*(6), 875–906. https://doi.org/10.1080/02671522.2020.1870543

Newton, P. E. (2011). A level pass rates and the enduring myth of norm-referencing. *Research Matters: A Cambridge Assessment Publication*, *2*, 7.

Norfolk, P. A., Farmer, R. L., Floyd, R. G., Woods, I. L., Hawkins, H. K., & Irby, S. M. (2015). Norm block sample sizes: A review of 17 individually administered intelligence tests. *Journal of Psychoeducational Assessment*, *33*(6), 544–554. https://doi.org/10.1177/0734282914562385

Nusche, D., Earl, L., Maxwell, W., & Shewbridge, C. (2011). *OECD reviews of evaluation and assessment in education: Norway*. OECD. https://www.oecd.org/norway/48632032.pdf

OECD. (2021). *Scotland's curriculum for excellence*. https://www.oecd-ilibrary.org/content/publication/bf624417-en

Ofqual. (2018). *Inter-subject comparability in A level sciences and modern foreign languages*. https://assets.publishing.service.gov.uk/media/5bf433ff40f0b60783ad9374/ISC_Decision_Document_20.11.18.pdf

Ofqual. (2022). *Teacher assessed grades in summer 2021: Interviews*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1089448/6933-2_TAGInterviewReport_110722.pdf

Ofqual. (2023). *Mathematics: Grade descriptors for GCSEs graded 9 to 1*. https://www.gov.uk/government/publications/grade-descriptors-for-gcses-graded-9-to-1/grade-descriptors-for-gcses-graded-9-to-1-mathematics

Ofqual. (2024). *Perceptions of A levels, GCSEs and other qualifications: Wave 22*. https://www.gov.uk/government/statistics/perceptions-of-a-levels-gcses-and-other-qualifications-wave-22

Ontario Ministry of Education. (2010). *Growing success: Assessment, evaluation, and reporting in Ontario schools*. Queen's Printer for Ontario. https://www.edu.gov.on.ca/eng/policyfunding/growSuccess.pdf

Opposs, D. (2016). Whatever happened to school-based assessment in England's GCSEs and A levels? *Perspectives in Education*, *34*(4), 4. https://doi.org/10.18820/2519593X/pie.v34i4.4

Opposs, D., & Gorgen, K. (2018). Chapter 4: What is standard setting? In J.-A. Baird, T. Isaacs, D. Opposs, & L. Grey (Eds.), *Examination standards: How measures and meanings differ around the world*. UCL, IOE Press.

Perryman, J., Ball, S., Maguire, M., & Braun, A. (2011). Life in the pressure cooker–school league tables and English and mathematics teachers' responses to accountability in a results-driven era. *British Journal of Educational Studies*, *59*(2), 179–195. https://doi.org/10.1080/00071005.2011.578568

Popham, W. J. (2001). Teaching to the test? *Helping All Students Achieve*, *58*(6), 16–20.

Popham, W. J. (2014). *Criterion-referenced measurement: Half a century wasted?* https://www.ascd.org/el/articles/criterion-referenced-measurement-half-a-century-wasted

Republic of Estonia Ministry of Education and Research. (2023). *Õppijate hindamiskriteeriumid—assessment criteria for learners*. Digipädevus. https://digipadevus.ee/oppija-digipadevusmudel/hindamiskriteeriumid/

Richmond, T. (2021). *Re-assessing the future: Part 1 – how to move beyond GCSEs*. EDSK. https://www.edsk.org/wp-content/uploads/2021/01/EDSK-Re-assessing-the-future-part-1.pdf

Rosen, M. (2017). Dear Ms. Greening, stop raising and lowering the GCSE high-jump bar. *The Guardian*. https://www.theguardian.com/education/2017/aug/22/gcse-grade-boundaries-altered-michael-rosen

Rust, C., & O'Donovan, B. (2025). Assuring academic standards: Policy and practice in context. In N. Reimann, I. Sadler, & J. Hill (Eds.), *Academic standards in higher education critical perspectives and practical strategies*. Routledge.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, *13*(2), 191–209. https://doi.org/10.1080/0305498870130207

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, *34*(2), 159–179. https://doi.org/10.1080/02602930801956059

Santori, D. (2020). Test-based accountability in England. In J. Lampert (Ed.), *Oxford Research Encyclopedia of Education*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264093.013.1454

SEC. (1984). *The development of grade-related criteria for the GCSE. A briefing paper for working groups*. Secondary Examinations Council.

Shepard, L. (1979). Norm-referenced vs. criterion-referenced tests. *Educational Horizons*, *58*(1), 26–32.

Singh, S. (2024). *Urgently adjust English language 2.0 GCSE grade boundaries set by Pearson Edexcel 2024*. https://www.change.org/p/urgently-adjust-english-language-2-0-gcse-grade-boundaries-set-by-pearson-edexcel-2024

Sizmur, S., & Sainsbury, M. (1997). Criterion referencing and the meaning of national curriculum assessment. *British Journal of Educational Studies*, *45*(2), 123–140.

Sternberg, R. J., Chowkase, A., Parra-Martinez, F. A., & Landy, J. (2022). Criterion-referenced assessment of intelligence as adaptation to the environment: Is it possible, plausible, or practical? *Journal of Intelligence*, *10*(3), 57. https://doi.org/10.3390/jintelligence10030057

Stringer, N. (2012). Setting and maintaining GCSE and GCE grading standards: The case for contextualised cohort-referencing. *Research Papers in Education*, *27*(5), 535–554. https://doi.org/10.1080/02671522.2011.580364

Taylor, R., & Opposs, D. (2018). Chapter 6: Standard setting in England: A levels. In J.-A. Baird, T. Isaacs, D. Opposs, & L. Grey (Eds.), *Examination standards: How measures and meanings differ around the world*. UCL, IOE Press.

The College Board. (2017). *SAT suite of assessments technical manual*. The College Board. https://satsuite.collegeboard.org/higher-ed-professionals/test-validity/test-development

Torrance, H. (2007). Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education: Principles, Policy & Practice*, *14*(3), 281–294. https://doi.org/10.1080/09695940701591867

Tveit, S. (2014). Educational assessment in Norway. *Assessment in Education: Principles, Policy & Practice*, *21*(2), 221–237. https://doi.org/10.1080/0969594X.2013.830079

Wikström, C., & Pantzare, A. L. (2018). Chapter 12: Setting standards in Sweden: School grades and national tests. In J.-A. Baird, T. Isaacs, D. Opposs, & L. Grey (Eds.), *Examination standards: How measures and meanings differ around the world*. UCL, IOE Press.

Wiliam, D. (1995). Combination, aggregation and reconciliation: Evidential and consequential bases. *Assessment in Education: Principles, Policy & Practice*, *2*(1), 53–74.

Wiliam, D. (1996). Meanings and consequences in standard setting. *Assessment in Education: Principles, Policy & Practice*, *3*(3), 287–308. https://doi.org/10.1080/0969594960030303

Wiliam, D. (2010). What counts as evidence of educational achievement? The role of constructs in the pursuit of equity in assessment. *Review of Research in Education*, *34*(1), 254–284. https://doi.org/10.3102/0091732X09351544

Wolf, A. (1995). *Competence-based assessment*. Open University Press.

Wright, B. D., & Bell, S. R. (1984). Item banks: What, why, how. *Journal of Educational Measurement*, *21*(4), 331–345.

How to cite this article: Meadows, M., Cadwallader, S., Gray, L., & Baird, J.-A. (2025). What do I have to do to get the grade? How examination standards interact with learning, teaching and the curriculum. *Review of Education*, *13*, e70126. https://doi.org/10.1002/rev3.70126