Standards in GCSEs in Wales: approaches to defining standards

Michelle Meadows, Jo-Anne Baird, Lena Gray, Stuart Cadwallader, Thomas Godfrey-Faussett, Luke Saville, Candace Debnam, & Gordon Stobart

September 2023

To cite this report:
Meadows, M., Baird, J., Gray, L., Cadwallader, S., Godfrey-Faussett, T., Saville, L., Debnam, C., and 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/

Table of Contents

Figures

Tables

Boxes

1 Foreword

We know there are lots of discussions happening across many jurisdictions about reforming assessment, including more use of digital technology, debates about exams and continuous assessment, and a myriad of other aspects of qualifications and assessments. Indeed we are engaging in these debates ourselves as we enter an era of change with new Made-for-Wales GCSEs in response to Curriculum for Wales.

What is less commonly focused on is the impact that reforms can have on the grades that young people receive - in terms of what those grades can be taken to mean by people that use them. This includes the young people themselves, as well as colleges and universities, employers and government at all levels.

Qualifications have important social functions, making the benefits of being educated more apparent to young people at the point of being educated. So we need to think through changes that could impact on grading and then how the qualifications are used.

This research is relevant to the commentary that we often hear about the grading of GCSEs, including that 'GCSE grading is norm referenced, but should be criterion referenced' and 'there is a fixed quota of GCSE grades'. It sets out to explain what these terms mean and makes clear that the approach taken to grading GCSEs is not, and never has been, norm referencing.

The report also sets out the challenges presented by a change in standards approach. Shifts in standards approaches are fraught with significant risks and a high likelihood of unintended consequences. However, the findings give us food for thought about improvements that could be made to the current approach.

Despite previous attempts, the qualification system has not been particularly successful in explaining how and why GCSE grading seeks to reflect the attainment of young people. The report helps us to understand why many factors contribute to the difficulty of clear communication.

So in commissioning this report we aimed to create the foundations needed for a shared understanding of how grading works and to start a discussion about how it can be more transparent.

Philip Blaker

Chief Executive

2 Executive summary

Purpose of the Standards in GCSEs in Wales project

1. This report is part of a broader project on standard setting in GCSEs in Wales that consisted of four linked strands of research. This strand includes a review of the standard setting literature to inform a description of the normal approach to standard setting, how standards are embedded throughout the qualification lifecycle and how a change of approach would affect assessment processes. The report also explores the advantages and disadvantages of alternative approaches, specifically norm- and criterion-referencing. As well as drawing on the academic literature, we conducted empirical research with policymakers, teachers, and other educationalists to investigate their views of standard setting for GCSEs in Wales and the ways in which standard setting has been communicated to stakeholders. This had the dual aim of enabling the validation of our description of the current standard setting approach and informing the production of future standard setting communications. The third strand of the project involved an investigation into the teaching of standard setting in postgraduate education and the production of teaching resources. The final strand of the project will outline some principles for an empirical study of criterion-referencing. Each of these activities forms the basis of separate outputs.

2. This research contributes to the debates on standards that are evident in Wales, as in other countries. We refer to the 'normal' approach to standard setting to distinguish it from methods used in extremis during the pandemic. Stakeholders need a good understanding of the normal approach to standard setting and a common language to discuss priorities for any future approach. We hope that the report provides increased clarity regarding approaches to standard setting, the implications of any change and definitions of terms that are often used in this area. In addition, we hope it supports high-quality discussions in the context of reforms to the curriculum and assessment in Wales. The purpose of this work is not to take a stance on how standards *should* be set for GCSEs in Wales, now or in future reformed qualifications.

Setting standards in GCSEs in Wales

3. Standard setting is the process of setting boundary marks that define which candidates' assessment performances are awarded a particular grade and which fall below expectations. The aim of standard setting is to ensure students are judged against consistent, defined standards, within and across cohorts, to support the intended use of grades. In common with other countries, the approach to standard setting has changed over time in Wales, especially during the pandemic, but this project looked at how GCSE standards are set in normal years in Wales. In keeping with England, Northern Ireland and Scotland, we categorise the approach to standard setting in Wales as **attainment-referencing** (see Table A 1).

Table A 1 Definitions of standard-setting approaches taken from the academic literature

	Candidates receive …
Attainment-referencing	grades that reflect their holistic attainment in the qualification at a standard which is comparable with the attainment required for that outcome in the previous years' qualifications
Norm-referencing	grades that tell us where they rank in relation to the population of students who could have taken the qualification in any year
Cohort-referencing	an outcome that tells us where they stand in relation to the population who took the qualification in the same series or year
Comparable outcomes	as a group, outcomes comparable to those which they would have received had they followed the course before a reform and taken the old qualification
Criterion-referencing	grades that tell us whether they met predetermined performance criteria

4. Although our focus is upon standard setting, we take a broad approach to describing GCSE standards in this report because standards must be embedded throughout qualification design, development, delivery and review. Threats to standards, and so to the validity of interpretations of GCSE grades, can arise from any of these aspects. Thus, we explain here how standards are embedded in the qualification lifecycle for GCSEs in Wales in normal years.

5. Standard setting itself can be defined in several ways. We take two main alternatives and explore how adopting each of them would affect not only the standard setting process for GCSEs, but the qualification lifecycle. These are important considerations when contemplating alternative approaches to standard setting; the broader implications and expectations that would follow need to be anticipated. **Norm-referencing** and **criterion-referencing** alternatives to attainment-referencing are explored most fully in this report, but we also discuss the implications of **cohort-referencing** and **comparable outcomes**.

6. GCSEs have multiple purposes and are part of the educational landscape in Wales, England and Northern Ireland. The outcomes are used to select students for further study and for the labour market. Some higher education institutions use them for selection to degree programmes. They are used as part of the accountability framework for schools. And, of course, they provide feedback to the students who have taken them about their attainment levels.

7. There are features of GCSEs in Wales that any standard-setting approach needs to manage. GCSE subjects are optional, so entry patterns change from year to year. Indeed,

there are alternative qualifications that students may take. Most GCSE entries are made at age 16 but it is possible for younger students to enter, for example at age 15, or older. English language and English literature, and mathematics and mathematics numeracy are often taught and assessed sequentially across years 10 and 11. Further, many GCSEs in Wales are modular (referred to as unitised) and entry patterns can shift from series to series.

8. Various bodies have responsibility for aspects of standards in GCSEs in Wales. The Welsh Joint Education Committee (WJEC), the main exam board for Wales, is responsible for the provision and management of GCSEs and their standards. They are regulated by Qualifications Wales. The Welsh Government sets the overarching policy for education, including the design of *Curriculum for Wales*. The inspectorate, Estyn, reports on the quality of education and training in Wales, including on providers that offer GCSEs. Additionally, over 200 schools and colleges have responsibility for the teaching and learning, some aspects of the assessment and elements of the administration of GCSEs.

9. Over 34,000 16-year-olds took GCSEs in Wales in 2022. Thirty-six subjects are available nationally. GCSEs in Wales are graded A* to G, whereas in England they are graded 9 to 1. There is more than one GCSE examination series in Wales. In the main summer series, results at grade C went up from just under 67% in 2017, to just under 75% in 2020, but they returned to just under 70% in 2022. This reflects changes to the assessment and standard setting processes during the pandemic.

How standards are embedded in the GCSE qualification lifecycle

10. GCSEs in Wales are regulated qualifications and can only be offered by exam boards that have been recognised to do so by Qualifications Wales. Conditions for recognition help to set an appropriate standard for the provider. Qualifications Wales also sets approval criteria and subject-specific criteria that control the assessment of the subject content, the assessment objectives and the structure of the qualifications. All of these have implications for assessment standards.

11. Demand of the assessments and their associated marking schemes, the administration of the assessments (including security of the question papers and conduct of coursework[1]), quality assurance of the marking, rules for special consideration and the appeals process all affects standards. Standard setting, through grade award meetings, is the specific procedure through which grade boundaries are set. How each of these procedures is

[1] Coursework is a form of non-exam assessment. We use the term coursework rather than non-exam assessment throughout the report to signify assessment conducted in conditions with lower levels of control than examinations and which is marked by teachers.

currently conducted and their relationship with standards is outlined broadly in this report. Changes to any detail or oversights regarding the execution of these processes can pose serious threats to standards.

12. Threats to GCSE standards are multifarious. Malpractice is a threat to standards under any definition of standards. Assessment formats affect the nature of the issues. Authenticating students' contribution to coursework has proved a major challenge, contributing to removal of coursework from GCSEs in many subjects in England. However, removing coursework has also been criticised, for undermining validity and negatively affecting teaching and learning, threatening standards in a different way. Questions over marking consistency have also been a controversial threat to standards at GCSE.

13. In attainment-referencing, a range of information is considered when the decision is being made about where the boundary marks should be set. This includes examiners' judgments of students' assessment performances and statistical information. It is therefore a 'mixed methods' approach. Since student performance can seem better or worse depending on the demand of the assessments, statistical information helps examiners account for the context in which the performances have been produced. This is a key strength of the approach. The weight that is placed on examiners' judgments versus statistical information varies according to their relative fidelity. Under some circumstances, examiners' judgments may be weaker than usual, for example when setting grade boundaries in qualifications with new subject content. Equally, statistical information can sometimes be weaker, such as when there has been a change in the motivation of the students taking the qualification.

14. Modular qualifications present particular challenges to any standard setting approach due to the complexities of early entry, re-sitting and variation in routes through the qualification. For these reasons, oversight of the final outcomes at GCSE can be more difficult to assure in modular systems, especially if the terminal modules (which must be taken at the end of the course) have a low weight in the overall grades. The multiple purposes and stakes of GCSEs place a great deal of pressure on standards throughout the qualification lifecycle.

The qualification lifecycle under norm-referencing

15. Norm-referencing would produce radical changes to how GCSEs were graded, as well as to their design and conduct. A separate norming study is first conducted when the assessments have been designed. This involves a representative sample of the population of interest, such as 16-year-olds in Wales who have prepared for the qualification, taking an unseen, secure test. Questions are not released publicly, because advance knowledge of them would affect their difficulty. This norming study would provide the basis for a

comparison of future students' scores with the population. For example, if a candidate had scored particularly highly, their score might be in the top ten percent of the population according to the norming study.

16. Under norm-referencing, in theory outcomes at a national level could rise or fall, as they would reflect how the sample of students taking the qualification performed compared to the population norm. However, in practice, outcomes tend to rise under norm-referencing as the education system becomes more familiar with the test questions, even when there are strong attempts to keep them secure. Inevitably, test-takers discuss items they remember with their teachers and other students.

17. Various assessment designs can be used to secure standards for each new assessment series, and these are underpinned by sophisticated psychometric, statistical procedures. However, coursework is unlikely to be compatible with norm-referencing because the assessments are normally held securely until they are conducted. Predictability is anathema to norm-referencing. Multiple-choice tests are often the format of choice for norm-referencing due to their suitability for machine scoring, the psychometric models and standardised administrative procedures.

18. Appeals would likely be limited to administrative checks, since access to the question papers would be closely guarded. Moreover, the relationship between assessment, teaching and learning would be influenced by this model, since the assessments are not openly published, as they are in the current model.

19. Threats to standards from norm-referencing pertain more to the effect upon teaching and learning than to the assessment systems themselves. GCSEs would look very different under such an approach – they would likely be multiple-choice tests and their content would not be public. This fundamentally affect the kind of learning that is anticipated and the associated pedagogy. Notwithstanding, specific threats to standards that arise from a norm-referenced system relate to the requirements for high security of the test, accuracy of the initial norming study and any statistical calibration required for the introduction of new items. It is perhaps for these reasons that we have not been able to identify a high stakes qualification system internationally that uses norm-referencing. Cohort-referencing, on the other hand, is used in several countries, such as Chile, South Korea, South Africa and Georgia.

20. The term norm-referencing is sometimes mistakenly used to describe cohort-referencing. **In cohort-referencing a fixed proportion of the cohort taking the qualification in a given year is awarded a particular grade.** This approach is most appropriate when taking the qualification is compulsory and the composition of the entry is similar annually. Cohort-

referencing has not been used as a standard setting approach in GCSEs in Wales, and the optional nature of GCSE entry would make it hard to defend.

The qualification lifecycle under criterion-referencing

21. A standards-referenced form of criterion-referencing is the basis of our consideration of the implications for standards in GCSEs in Wales, since it is closest to the assessment culture now and paradigm shifts in standard setting are exceptions internationally.
    Policy positions on several matters had to be assumed for the purposes of this exploration, such as
    a. text-based criteria devised by subject experts being publicly available
    b. holistic performance judgments being made using these criteria, over a series of assessment tasks
    c. more detail than in the current assessment criteria (but not atomistic)
    d. teachers/assessors would be provided with the appropriate tools and training to conduct the assessments, as part of a community of practice
    e. moderation procedures would be put in place to assure consistency of grading

22. Under criterion-referencing, a great deal of emphasis is placed upon the development of the criteria, including the credentials of the people involved in their specification. Experience shows that it is impossible to create criteria that will be understood in the same way by all assessors. Attempts to do so may lead to an unmanageable number of narrow criteria that are still open to some interpretation and that drive poor pedagogy and assessment burden. The inherent imprecision of the language used to describe the criteria means that policymakers must accept that exactness, and therefore strict criterion-referencing, is not, in fact, desirable or possible.

23. Teacher assessment formats are typically seen as more compatible with a criterion-referencing ethos, though examinations can be used. More emphasis is placed upon professional development of teachers, who are more likely to be assessors, with direct responsibility for grading, under this model. Checks on standards often relate to verifications that the procedures have been followed under criterion-referencing. Appeals would also likely involve checks on procedures. Since teacher assessment is the norm, appeal procedures may begin in the school or college but would then likely involve processes external to the centre.

24. Under a criterion-referencing approach, in theory it would be acceptable for all candidates (or no candidates) to pass. Everything depends upon whether the candidates are judged to have met the published criteria. In practice, results have been found to vary dramatically unless there is statistical input to standard setting. This is achieved in a variety of ways under different systems. For example, there might be an assumption that

the results should be normally distributed and therefore assessors, in effect, grade on the curve.

25. Many of the threats to standards under a criterion-referencing system of setting standards emanate from its distributed, decentralised nature. Empirical evidence and documented operational experience with criterion-referencing has shown that consistency of judgments can be more problematical than with centralised marking and grading. Nonetheless, some high stakes qualification systems use criterion-referencing, but there are different expectations about the stability and comparability of outcomes compared with GCSEs for those qualifications. For example, some vocational qualifications in the UK adopt this approach, such as NVQs. Additionally, school-leaving assessments in France, Sweden and Queensland are criterion-referenced. Appropriate moderation systems would be required. One corollary of criterion-referencing that would have to be closely managed is teacher workload, which would increase.

26. Outcomes of criterion-referenced assessments, like the grading criteria themselves, are sometimes atomised, with profiles rather than an overall grade. This can be a threat to the coherent signalling of an overall standard of the qualification. Consistent learner support is a threat to standards, since feedback until the criteria have been met is part of the pedagogical process, but this is unlikely to be standardised across the education system. Malpractice has its own particular context in such a distributed system, posing threats to standards. Monitoring and controlling malpractice have to be part of the professional work of teachers, as well as the exam board. Accountability regimes can threaten standards in systems that depend upon teacher professionalism, since they can set up incentives and conflicts of interest.

Implications of Curriculum for Wales for standard setting

27. Qualifications are reformed regularly in most countries to meet the changing demands of society. *Curriculum for Wales* is the most recent set of national reforms. Changes to the design of qualifications have consequences for standard setting.

28. GCSEs aligned with the reforms are currently being developed. Their design has been informed by a number of principles, including the desire to promote well-being and mental health, and positive teaching and learning experiences. They are mostly modular and there is more coursework than in the current GCSEs or the GCSEs in England. Data on GCSE outcomes will be used, as part of a wider set of information, for evaluation and accountability purposes.

29. Decisions about the reformed GCSEs designed to align with the *Curriculum for Wales* need to be informed by the latest technological developments. Recent advances in generative

artificial intelligence are likely to change the way in which students and teachers work. Such technology can automatically generate sophisticated long-form assessment responses that are difficult to reliably differentiate from those authored by students. These advances are likely to have significant implications for assessment, particularly standards in coursework.

30. How standards will be set for the reformed GCSEs has yet to be decided. Neither strong criterion-referencing nor norm-referencing are ideal standard-setting approaches for the new GCSEs. Norm-referencing is not a good fit, because, among other reasons, not all of the subject content can be validly assessed by examination. A move to criterion-referencing would require an (unlikely) acceptance that the standards applied across schools and colleges, and therefore across students, would be somewhat inconsistent, whatever the investment in the development of criteria and teacher development. This could manifest in fluctuations in national outcomes which would be difficult to investigate and to defend. In systems that employ criterion-referencing it is accepted that the standards/outcomes may be inconsistent. Moreover, **major paradigm shifts in standard setting are rare internationally and have inherent risks, since stakeholders have long-standing expectations regarding how qualifications operate.** Fundamentally changing the approach to standard setting would impact many of the current assessment processes and features of the GCSE, the consequences of which would require significant piloting and communication.

31. A comparable outcomes approach has been used to steady the system in recent reforms. In Wales, this involved increased weight being placed on statistics to ensure that outcomes are broadly comparable to those which the cohort would have received had they followed the course before a reform and taken the old qualification. However, the increase in coursework for the reformed GCSEs and their modular structure may present challenges to this approach, for example in the form of compressed grade boundaries on the written examinations. Attainment-referencing would face the same problems. The separate reporting of grades for non-examination and examination assessment would alleviate these issues but produce more complex information for qualification users regarding the overall standard of the qualification.

32. Modular assessment also presents challenges to cohort-referencing and attainment-referenced methods. Some systems report only unit level standards for this reason, which removes the difficulty of ensuring comparable standards across all possible routes through the qualification but can create obscure grading processes for users of the grades.

33. Criterion-referencing is designed for a distributed, teacher-led assessment system. As we have seen, there are challenges to standards using this approach. These would equally apply to the *Curriculum for Wales* GCSEs. Designing a moderation system to support

consistency in grading standards would be important if this approach were to be adopted. This would likely include the adoption of ongoing social moderation rather than moderation by inspection, which is currently used in Wales. Even with a system of continuing social moderation, it is likely that some inconsistencies in judgments would remain.

34. Rather than the wholesale adoption of criterion-referencing, there may be elements of the approach that could be adopted to support the content meaning of the grades and give clarity to the curriculum aims being assessed to allow teachers to teach better. Despite the shortcomings of grade descriptors, teachers may well find them useful alongside other exemplification of the standard required for each grade, such as candidate work with related commentaries. There may also be opportunities to build a stronger community of practice around assessment standards by drawing on elements of social moderation common in criterion-referenced systems. This may help build teacher understanding of assessment in the context of *Curriculum for Wales*.

35. The geographical proximity of England and Wales creates permeable boundaries between the two education systems. Given this and the use of the same qualification title, there will be an expectation that standards in GCSEs in Wales are comparable with those in England. Given the increasing divergence of the content, structure and assessment of GCSEs in the two countries, perhaps the most appropriate interpretation of this expectation is that students achieving a particular grade are equally well prepared for future study or training, whatever the country in which they were awarded the qualification. Any evaluation of potential standard setting options might consider the implications of this expectation.

36. Post-pandemic, *Curriculum for Wales* represents an opportunity to revisit the approach taken to setting standards for GCSEs in Wales and to build a broad understanding of standards within the teaching profession. An open dialogue, involving all stakeholders in the education system, of the implications of any of the possible approaches to standard setting is needed. This report seeks to inform that discussion.

3 Introduction

This project focuses upon how standards are set for GCSEs in Wales. Assessment standard setting is typically considered as the process of setting pass marks, otherwise known as grade boundaries, or cut-scores. This is often done when there is data on how candidates have performed in an assessment, and this is certainly the case for GCSEs. We refer to this as the *performance standard*. Different standard setting procedures have been documented and researched internationally (e.g. Baird et al., 2018; Brennan, 2006; Cizek & Bunch, 2007). Each national assessment has its own ways of approaching standard setting, since school-level qualifications are embedded in the context of the assessment and education systems and in the country's culture more widely (Isaacs & Gorgen, 2018). As such, documenting the procedures publicly is important. Indeed, a lot of effort has gone into making standard setting procedures public, evaluating them and to comparing them with other systems to improve them. In so doing, the assessment community has had to formulate terms to define assessment concepts and procedures. This is necessary in all fields, or we cannot build upon previous work. However, the technical terms, statistical procedures and acronyms that are used have also made it more difficult for the assessment community to communicate easily with stakeholders. Miscommunication and misunderstandings are common. This project seeks to contribute to reducing the communication gap.

Although the project focuses upon GCSE standard setting in normal (non-pandemic) times, there were clear lessons to be drawn from the pandemic years. One communication gap that became patently obvious during the pandemic was that stakeholders did not view qualification standards as being solely the province of the standard setting process. What was required of candidates was also viewed as being part of the standard of a qualification, which of course it is. In the assessment community, we refer to this as *content standards*, which outline the knowledge, skills and understanding that learners should acquire to gain the qualification. Typically, this is communicated through a syllabus (specification), though the question papers are also communication devices which shape teachers' and students' understandings of the content standards.

Previous work on assessment standards has made the distinction between performance and content standards (e.g. Opposs & Gorgen, 2018), but the way in which content standards are embedded in qualifications has not been conceptualised as comprising the standard of a qualification. Naturally, though, the design of the question papers and the marking scheme, how the latter is applied, the training and quality assurance of the marking process, the nature of the appeals system and so on are crucial to upholding the standard of a qualification. Changes to the approach to setting standards affect the entire lifecycle. This report explains how each stage in a qualification lifecycle contributes to the setting and upholding of standards in the system for GCSEs in Wales, in normal times. We also consider what the

qualification lifecycle and associated assessment processes would look like under different methods of setting standards – norm- and criterion-referencing. We discuss the benefits and tensions of each approach. At the system level, there are effects on outcomes which will have implications for how grades can be used.

The current research does not extend the discussion to A-levels in Wales, but as the processes for assuring the standards for A-levels are similar to those used for GCSEs, the findings can be extrapolated. Further, the current qualifications reform agenda in Wales implies particular methods for standard setting, and we consider this too.

While a great deal of research and policy work has been carried out in the UK on the meaning and communication of standards in national qualifications, Wales, because it has a devolved government, has a distinctive approach. The current project is therefore important in codifying the current Welsh approach to standards in national general qualifications in technical terms that would be recognised in the assessment field. As well as this report, this project will work on the communication of standards in Wales, how the topic of standards is taught in higher education and consider how empirical studies of the quality assurance of criterion-referenced teacher assessment could best inform future work. We are fortunate to have the advice and support of Qualifications Wales and WJEC, as well as the Oxford University Advisory Group (Appendix 1). Notwithstanding, this research has been conducted independently and the views expressed are the authors' own. As Wales is currently developing reformed qualifications which potentially have significant implications for the way in which standards are set and maintained, this project is timely.

The primary aim of this research is to describe how GCSE standard setting in Wales operates currently and the implications of adopting either a norm-referenced or a criterion-referenced approach. To inform our analysis we conducted a review of the standard setting literature and interviewed industry insiders.

3.1 Method

3.1.1 Review of literature

The focus of the review was articles and books relating to standard setting approaches, their history, their pros and cons, and the qualification or testing systems in which they have been applied. Search terms included standards setting, standards maintaining, norm-referencing, cohort-referencing, attainment-referencing, comparable outcomes and criterion-referencing. We were conscious that many relevant papers may be 'grey literature' – unpublished or published on exam board and regulator websites – so we contacted the four exam boards in England (AQA, OCR, Pearson and WJEC) and the regulators of the devolved administrations (CCEA, Ofqual and Qualification Wales) to request such papers. In addition,

WJEC provided documents setting out their approach to GCSE standard setting, which enabled us to check our understanding.

3.1.2 Interviews

3.1.2.1 Participants

We asked Qualifications Wales and WJEC were approached to suggest potential participants who would be able to inform the project with regard to current assessment procedures and to comment on their views regarding standards and their communication. Interviews were conducted across these organisations, including with participants with expertise in policy-making, assessment design, marking and appeals procedures, and standard setting. Between six and twelve interviews were anticipated when the project was planned. Nine interviews with participants from the organisations, involving ten participants were conducted.

3.1.2.2 Procedures

Interviews were mainly held on Microsoft Teams (n=9). Participants were sent the consent form and an information sheet about the project. Interviewees were also sent the following definitions in advance of the interviews:

Norm-referencing - Candidates receive grades that tell us where they rank in relation to the population of students who could have taken the qualification in any year. A formal norming study is required to understand the ranks for the broader population who could have taken the qualification.
Cohort-referencing - Candidates receive grades that tell us where they rank in relation to the population who took the qualification in the same year. (Based on Wiliam, 1996)
Attainment-referencing – Candidates receive grades that reflect their holistic attainment in the qualification at a standard which is comparable with the attainment required for that outcome in previous years' qualifications. (Based on Newton, 2011)
Comparable outcomes – Candidates receive, as a group, comparable grades to those which they would have received had they followed the course before a reform and taken the old qualification. (Based on Cresswell, 2003)
Criterion-referencing – Candidates receive grades that tell us whether they met predetermined performance criteria. (Based on Popham and Husek, 1969)

Interviews were semi-structured and focused upon the areas of expertise that the participant brought to the project. The participants were therefore asked specifically about their areas of responsibility, as well as their wider views. Each interview lasted between half an hour and an hour. Broad questions on the meaning of standards and fairness were pursued. Participants were asked to describe GCSE standard setting in their own words. Knowledge of terms for standard setting was probed (e.g. criterion-referencing, norm-referencing, attainment-referencing). Where participants were knowledgeable about these terms, a

discussion on their perceptions of the advantages and disadvantages followed. The need for comparability with GCSEs in England and Northern Ireland was explored. Opportunities for clarification regard elements of the assessment process were taken where appropriate. Finally, participants' beliefs regarding the implications for standards of the curricular reforms arising from the *Curriculum for Wales* policies were probed.

3.1.2.3 Analytical strategy

Interviews were recorded on Microsoft Teams and a transcript of each discussion was generated by the software. Structured notes were taken by a member of the project team during the interviews and these provided the basis for the initial data analysis. All of the interviews were listened to while the transcripts were read. The purpose of the interviews for this report was to establish a clear understanding of the current approach to setting standards at GCSE and to identify pros and cons of other approaches identified in the research literature. The interviews informed and validated our description of the current standard setting approach and informed the production of standard setting communications in another phase of the project.

3.1.3 Ethics

This study was conducted in accordance with the guidelines of the British Educational Research Association (BERA). Before data collection commenced, research ethics approvals were obtained for the study through the University of Oxford's (ED-CIA-2223-081) ethical approval procedures. All participants were sent a consent form and participant information sheet, which outlined how the data for the project would be stored and used. Consent was given, either through completion of the consent form or via oral consent as part of the interview.

Rather than including quotations from interviewees, the content of the interviews was incorporated into the report as appropriate, for example in the descriptions of the qualification lifecycle and the impact of a change of standard setting approach.

3.2 GCSE and A-level qualifications in Wales

> *GCSEs (General Certificates of Secondary Education) are available in a range of subjects. They are the main general qualifications taken by 16-year-old learners in Wales. They can be used as a basis for further study or training, or direct entry into employment.*
>
> (Qualifications Wales, 2023a)

GCSEs have multiple purposes. They indicate a basis for progress to further study and training and a signal to employers regarding the attainment level of applicants in the labour market. Qualifications Wales' (2019, 2021) regulations set out that the current GCSEs must:

1. provide evidence of learners' achievement against challenging and relevant content;
2. allow learners to develop a strong foundation of knowledge and skills which will support further academic and vocational study, as well as employment;
3. provide suitable preparation for learners, to enable them to progress to a GCE AS or A-Level in the same, or related, subject;
4. where appropriate, support opportunities to develop skills that are being assessed through the Welsh Baccalaureate

Approval criteria for the new suite of Made-for Wales GCSEs arising from the *Curriculum for Wales* policy set out the purposes and aims as follows,

- be designed primarily for learners between the ages of 14 and 16
- build on the conceptual understanding learners have developed through their learning from ages 3-14
- support teaching and learning by providing appropriately broad, demanding, relevant and engaging content and assessment that relates to and supports the Curriculum, including its four purposes
- allow learners to develop a strong foundation of knowledge, skills and understanding which supports progression to post-16 study and prepare them for life, learning and work
- provide meaningful, fair and accurate information on learner achievement within a subject that highlights what learners know, understand and can do

Qualifications Wales (2023b, p.18)

AS and A-levels are the main general qualifications usually taken at ages 17 and 18 respectively. A-levels are used as a basis for admissions to higher education, further training or entry into employment. Organisations responsible for relevant aspects of the education and assessment system in Wales can be found in Appendix B.

GCSE and A-levels are a common brand with England and Northern Ireland, though the Welsh qualifications have distinctive features. Devolved government policy decisions led to different structures, with Wales and Northern Ireland offering some modular GCSEs, while England moved to a linear assessment structure. In Wales, outcomes at AS level contribute to overall A-level grades, which is not the case in England, where the AS is a stand-alone qualification. Welsh schools receive funding for entry for approved GCSEs through the examination board, WJEC. For subjects not offered by WJEC, state schools can gain funding for entry to assessments offered by the examination boards in England and Northern Ireland.

The age 16 cohort in 2022 was 34,365 (Office for National Statistics, 2022). In 2022, 318,590 GCSE entries were made in 36 subjects and there were 36,310 A-level entries in 36 subjects (Qualifications Wales, 2022b). There werer 205 maintained secondary and middle schools (StatsWales, 2022) and around 11,066 secondary school teachers in Wales (Welsh Government, 2022).

At both GCSE (Table 1) and A-level (Table 2), outcomes were higher during pandemic years. For some this represents grade inflation, since pupils were not able to access the same levels of schooling as in normal times. Others considered that the teacher assessments allowed pupils to demonstrate their knowledge, skills and understanding more readily. This illustrates different perspectives on what it means to set standards. However, it is not the purpose of this report to address the comparability of standards in GCSE over time. Those interested in that topic can consult the findings of the National Reference Test, introduced by Ofqual for the purpose of monitoring standards at GCSE in England (Burge & Benson, 2022).

Table 1 Overall GCSE grade distribution in summer 2016–2022, 16-year-olds only[2]

Entries		A*	A	B	C	D	E	F	G
2016	218,232	6.8	21.4	42.8	69.6	85.3	92.7	96.8	98.9
2017	234,029	6.8	20.1	41.3	66.7	81.3	90.1	94.4	97.3
2018	242,034	6.2	18.6	38.1	62.8	78.4	87.6	92.8	96.5
2019	261,001	6.3	18.6	38.7	63.8	78.6	87.9	93.4	97.3
2020	263,881	11.3	25.9	48.8	74.6	87.3	93.8	97.2	99.6
2021	282,831	14.2	29.5	51.8	74.4	85.9	92.6	96.0	98.6
2022	266,552	11.4	25.8	46.5	69.7	82.1	89.7	94.3	97.5

Table 2 Overall A-level distribution in summer 2016–2022, all candidates

Entries		A*	A	B	C	D	E
2017	33,294	8.3	25.0	50.1	75.3	90.9	97.7
2018	32,445	8.7	26.3	52.0	76.3	91.0	97.4
2019	32,320	8.9	26.5	52.0	76.3	91.3	97.6
2020	30,513	16.7	42.3	70.7	91.8	98.7	99.9
2021	35,867	21.3	48.3	73.0	89.2	95.9	99.1
2022	35,499	17.1	40.9	66.5	85.3	94.3	98.0

Source: Joint Council for Qualifications website[3]

[2] There is also a smaller November examination series for GCSE.

[3] Examination results - JCQ Joint Council for Qualifications

Adjustments to assessments were made during the pandemic, with 2022 being the first year that examinations had been taken since 2019 (Table 3). In 2020, examinations were cancelled, and teachers judged what grade their students were likely to have gained had they continued with their studies in school rather than suffering the disruption that school closures from March 2020 caused. Evidence for grading was taken from a variety of sources, such as coursework, other assessments conducted in school and formative assessments of students in class. The advent of the pandemic meant that *post hoc* adaptations had to be made rather than there being systematic, planned data gathering to inform the assessments.

In 2021, with more time to prepare, more formal systems were established to gather performance evidence for the teacher-judged grades. As there was a great deal of disruption to schooling in the 2020–2021 academic year, adaptations were made so that students would only be assessed on the syllabus material they had been taught. In effect, a criterion-referenced approach to standard setting was used in 2021. We explain criterion-referencing in the next section.

In 2022, examinations were reintroduced, but subject-specific provisions were made to allow for the ongoing disruption to students' schooling experiences. For example, pupils only had to complete two (rather than three) units in English literature and three (instead of four) units in history. There was also an adjustment to grading standards, the intention being that outcomes would broadly fall midway between those from 2021 and those from 2019. Stronger use of statistics in this manner made the standard setting process more akin to cohort-referencing in 2022. We explain cohort-referencing in the next section.

Table 3 Assessment arrangements during the pandemic

Year	Examinations	Centre Assessed Grades	Performance evidence required	Adaptations to assessments
2019	✓		✓	
2020		✓		
2021			✓	✓
2022	✓		✓	✓

3.2.1 Standard setting

The method used to set standards and the meaning of such standards varies across assessments. Inferences that can legitimately be drawn from results will differ depending on the approach to standards that has been taken. We have grouped the approaches based on the type of evidence primarily used.

3.2.1.1 Approaches to standards based upon statistical information

Pure statistical methods do not tell us on their own what candidates have had to do to gain the result they were awarded. That information must be inferred from the other processes for embedding standards in qualifications. Here, we explain three statistical methods: norm-referencing, cohort-referencing and comparable outcomes. In fact, the way in which the comparable outcomes method was operationalised in Wales was not purely statistical; in terms of the academic literature, it was more of a mixed-methods, attainment-referencing approach. We outline it here in a way that is in keeping with how the comparable outcomes method is defined in the academic literature.

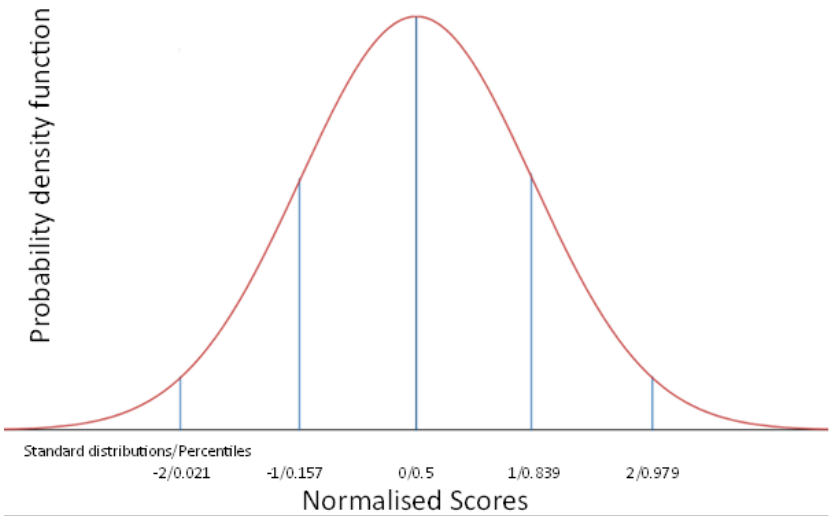
3.2.1.1.1 Norm-referencing

Definition Candidates receive grades that tell us where they rank in relation to the population of students who could have taken the qualification in any year.

Method To construct norms for an assessment, carefully conducted studies are carried out on representative samples of the population of interest. For example, intelligence tests can be normed for adult populations by testing a representative sample of the population with the tests and deriving what proportion score at each level. The tests are constructed so that the scores have a normal, or bell-curved distribution (Figure 1); and if the raw scores do not lie in this distribution, the raw scores can be mathematically transformed into scores which do follow a bell curve. This is useful because we know the statistical properties of a bell curve: the mean is in the middle of the curve with equal numbers of students scoring higher or lower. Approximately two thirds of students' scores lie within one standard deviation of the mean and 95% within two standard deviations, and we know the cumulative percentage of scores that will lie within a certain standard deviation from the mean. Armed with the knowledge of the population mean and standard deviation and the fact that the ability in question is normally distributed in the population, strong conclusions can be drawn about an individual's test score.

Interpretation Once a norming study has been conducted robustly, we can compare an individual's score on the same test with the distribution of scores for the general population and can draw conclusions about whether their test score is above or below average, or indeed what percentage of the population would score up to that level.

Figure 1 Distribution of a norm-referenced test in the population of the norming study



Origin Norm-referencing derives from psychological testing in the early part of the 20th century. Although Stern first used the term intelligence quotient, the 1908 Binet–Simon intelligence test was the first to group tests into age levels, assigning test-takers a mental age. Terman extended this approach to adult testing in 1916 (Boake, 2002). Intelligence tests such as the widely used Weschler Adult Intelligence Scale are designed to have a mean score of 100 in the population, with a normal, bell-curved distribution and a standard deviation of 15.

Main strengths and problems Since such strong inferences can be made from this approach, it has its attractions. However, it is based upon some strong assumptions. One such assumption is that the scores on the test are not affected by prior knowledge of the content of the test. Most tests using norm-referencing are kept secure, with test-takers not being able to access them for practice purposes, even if similar items are made available. Calls for transparency in national assessments have meant that the question papers are often published openly after tests have been taken, rendering them unusable in this way in future years. Additionally, the norming studies that would be required for a norm-referenced approach to assessment would be very costly for the multiple GCSEs (and other assessments) that are made available in a Welsh education context. As the tests themselves remain static under this method, standards are expected to be the same every year. Reductions in the relevance of the tests over time and public exposure of the test items undermine this assumption. In practice, the content of GCSE curricula is updated frequently.

In theory, results could rise or go down in any year, using this method. In fact, intelligence tests, which use this method, have seen rises in outcomes (Flynn, 1987; Trahan et al., 2014), with explanations for this effect ranging from increasing genetic variation due to more random mating patterns, socio-environmental improvements in living conditions and diet,

changes in the early environment (parenting styles, childcare and family size, education, increased abstract thinking requirements in society) and specific measurement effects (e.g. changes to the verbal ability scales may have counteracted the effect). We further discuss norm-referencing in relation to general qualifications later in this report.

3.2.1.1.2 Cohort-referencing

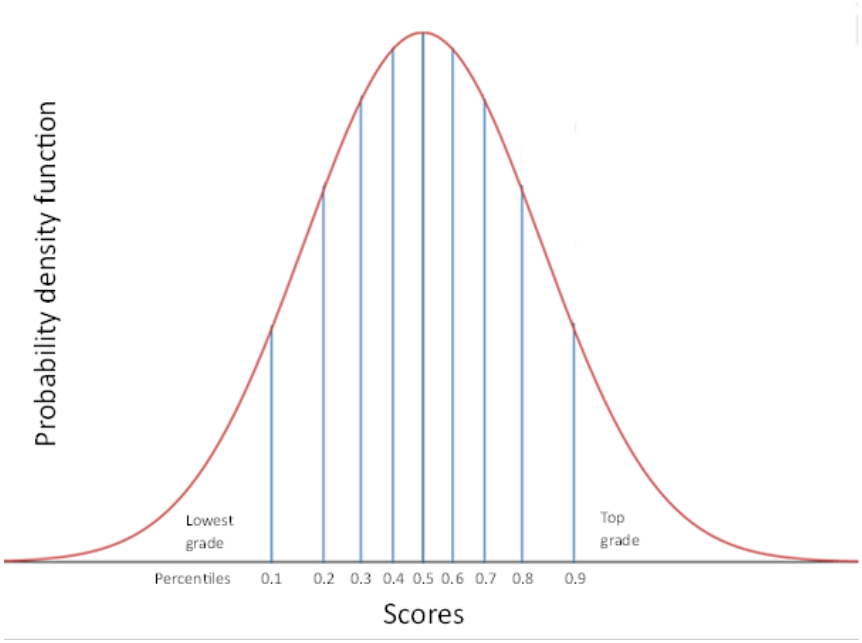
Definition Candidates receive an outcome that tells us where they stand in relation to the cohort who took the qualification in the same series or year.

Method GCSE question papers are only kept secure until the exam is held; after they have been used, they are made public. Once they have been released in this way, the difficulty of the questions would change because people would be able to practise for them and there could be teaching (directly) to the test. As such, norm-referencing in its pure form is not feasible. Cohort-referencing, in which that year's results are ranked without pre-testing the results on a representative sample of the population, however, is possible. A fixed proportion of students are awarded each grade each year under this model. For example, 10% of candidates might be awarded the top grade each year, with the next 20% being awarded the next grade and so on (Figure 2). Under cohort-referencing, the distribution of scores may not be normally distributed, as illustrated in Figure 2; a non-normal distribution is compatible with cohort-referencing.

Interpretation With cohort-referencing, we can interpret the results as telling us what the candidate's position was in relation to others who took the test in the same sitting. These are usually test-takers in the same age group. We cannot infer how this compares with candidates from other years. As with norm-referencing, percentiles can be used to reflect the outcomes. Equally, in either method, the outcomes can be grades.

Origin Wiliam (1996) pointed out that standards were never norm-referenced for national qualifications such as A-levels because no norming studies were conducted. When results were compared in terms of percentages, the comparison was with the specific cohort taking the test across years. Thus, he coined the term cohort-referenced.

Figure 2 Cohort-referencing



Main strengths and problems This method of setting standards is relatively simple and cost-effective and not a great deal of statistical expertise is needed. It requires no pre-testing. With the level of choice available to candidates at GCSE, there are problems in interpreting the results over time. GCSE English and mathematics are taken by the majority of the population, so for those subjects, we might reasonably infer that the relative standing of the candidate with respect to the population of test-takers is a comparison of the candidate with virtually all students for a given year. However, even for English and mathematics there would be complications if, for example, candidates from other age groups were to take the qualifications. The level of variability in the subject choices between years make interpretation even more difficult in other subject areas. There are circumstances in which this method could lead to indefensible grade boundaries. For example, GCSE grade G boundaries could be so low as to undermine confidence in the expected standard.

3.2.1.1.3 Comparable outcomes

Definition Candidates receive, as a group, comparable grade outcomes to those which they would have received had they followed the course before a reform and taken the old qualification.

Method A statistical prediction is made for the group of candidates taking the qualification. This is done using prior attainment data if it is available, for example in A-level standard setting prior attainment at GCSE is used. At GCSE no prior attainment data is available. The statistical evidence used to grade GCSE in Wales takes the form of the outcomes for centres that are common to the reference year and the current year in a subject. The assumption is that the outcomes for these common centres will be stable from one year

to the next once aggregated across centres. On this basis, results are predicted for the common centres. The predictions guide the setting of boundaries, which are then applied to all students.[4]

Interpretation The rationale for comparable outcomes is that the population of candidates taking the qualification should be awarded a similar profile of grades to those who took the qualification in the past, if the general academic ability of the students entering the qualification and the kind of centres teaching the qualification have not changed. Cohorts of pupils taking the qualification in years when the qualifications are newly revised, or when there are significant disruptions of other kinds (e.g. a pandemic, teacher strikes or school closures) are not disadvantaged compared to cohorts in other years.

Origin The term 'comparable outcomes' was devised by Cresswell (2003) to address the effects of curricular and assessment reform upon outcomes, that is, to make comparisons between one cohort and the next fairer. The idea was to smooth out the Sawtooth Effect: an initial downward turn in performance, followed by rising outcomes which were known to follow reforms due to the system's lack of familiarity with the new assessments (Linn, 1995). The use of this kind of approach outside of a period of qualification reform or disruption has been termed contextualised cohort-referencing (Stringer, 2012).

Main strengths and problems Comparable outcomes is beneficial when managing standards at a systems level, since a relatively stable level of grading results from this process. This means that societal systems which use grades as an indicator (e.g. further education, higher education and employers) can reasonably expect a predictable number of people will gain the grades. Additionally, those individuals taking the qualification in a given year benefit from the predictability at a systems level because of the stability in the distribution of grades. Comparable outcomes can be controversial because of its failure to incorporate information about changes in performance, for example due to school improvement or decline. The method has been criticised for capping student outcomes (e.g. Blatchford, 2020). Like cohort-referencing, there are circumstances in which this method could lead to indefensible grade boundaries. Anticipating this, Cresswell (2003) argued that for ethical reasons, in times of reform, where there is conflict between maintaining comparable performance and comparable outcomes, the use of examiner judgments of performance should only prevail

> *where the requirements of the comparable outcomes perspective cannot be followed through completely without serious damage to the credibility of the examinations.* (p. 16)

[4] The term comparable outcomes has sometimes been used by policymakers to describe approaches that would be termed attainment-referencing in the academic literature because they include an element of examiner judgment as well as statistical evidence.

Cresswell (2003) also argued that the statistics might need to be the dominant source of evidence for a number of years after the reform, while performance improves. The use of statistical predictions makes strong assumptions regarding the stable nature of the cohort – that the same kind of students are entered each year – which are unlikely to be met perfectly.

3.2.1.2 Approaches to standards based upon qualitative judgments

3.2.1.2.1 Criterion-referencing

Definition Candidates receive grades that tell us whether they met predetermined performance criteria.

Method Suitably qualified subject matter experts define the performance criteria required for the award of a qualification and/or for particular grades. In GCSE, grade descriptors could be thought of as the performance criteria to be met. These also relate to attainment objectives. Standard setting is not, strictly, a distinct phase in a system in which criterion-referencing is used, because the grades can be allocated directly by qualified assessors. Instead of standard setting, verification processes are applied to quality assure that all due procedures have been heeded in the assessment process and that an appropriate standard has been applied. In some cases, moderation is used, in which case the assessment judgment is also quality assured.

Interpretation The usual interpretation of outcomes of a criterion-referenced assessment is that students have demonstrated the required performances for the grade to be awarded. In the case of GCSEs, this is the knowledge, understanding and skills in attainment objectives and the specification. Thus, the idea is that absolute standards regarding performances can be inferred from the grades awarded – for example, the understanding of mathematics would be at the same level for a grade C for any given year in which the criterion-referenced assessment was awarded. This approach is used widely in vocational and technical qualifications and in higher education.

Origin Although criteria for assessments have in all likelihood been in use for many centuries, the term criterion-referenced testing was first used by Glaser (1963) to distinguish assessment of 'absolute' performance standards from relative ranks derived from norm-referenced tests.

Main strengths and problems Criterion-referencing is, on the face of it, a simple way to grade. It is an approach that recognises the professionalism of the assessors; teachers in the case of GCSE. It also addresses what students have to do rather than simply comparing them with each other. Therefore, in theory, everyone could pass, or no-one might pass the qualification. The onus in criterion-referencing assessment lies with the assessors, to identify the standards in writing, produce assessments of the required (and stable) levels of demand

and to assess consistently to the right standards. In practice, assessments vary in demand unless there is a pre-testing trial. Additionally, as with marking, understanding and application of the criteria vary between assessors. Instability in outcomes for national qualifications arises when they are criterion-referenced (Baird, 2007). There are also significant workload implications for assessors (teachers) in such a system. We discuss criterion-referencing further in relation to general qualifications in a later section of this report.

An example of a criterion-referenced qualification is the SQA (Scottish Qualifications Authority) National 4 Unit in Numeracy (Box 1). It is one of four units of the National 4 (broadly speaking, equivalent to GCSE Foundation Tier) Course in Applications of Mathematics.

3.2.1.3 Approaches to standards based upon mixed methods

Most national assessment systems use a combination of qualitative judgments and statistical methods in standard setting (Cizek & Bunch, 2007, p. 10). For example, statistical evidence about the cohort and their performance may be used alongside expert qualitative judgments about assessment demand or typical performance observed around particular grades boundaries. For this reason, Opposs and Gorgen (2018) refer to standard setting as a mixed-methods informed decision, with differences between the procedures relating to how information is integrated:

- Is the standard setting policy deductive (driven by a theory e.g., that outcomes will be similar year on year) or inductive (driven by observations e.g., of the quality of exam responses)? Quantitative evidence usually dominates where policy is deductive – where there are firm reasons to expect a particular outcome distribution. This may be the case when a qualification has been in place for some time.
- What data are collected for the standard setting process? Depending upon the view of standard setting, different kinds of evidence are collated for the decision-making process. Statistical information about performances or question difficulty may be seen as more, or less, pertinent than qualitative information from students' performances. Qualitative and quantitative information can be core or supplemental and there may be multiple forms of each.
- In what order are the components of information collected and presented?
- When are the sources of information combined? This can happen as the data are being analysed, or at the point of decision making. Different people may be involved in each stage.

Decisions regarding each of the above questions, among others, can produce standard setting systems that appear quite distinctive, even if they use the same definition of standards.

Box 1 Excerpt from SQA National 4 Unit in Numeracy

Outcome 1 The learner will:

1 Use reasoning skills and financial skills linked to straightforward real-life contexts by:

1.1 Interpreting a situation involving finance and identifying an appropriate strategy

1.2 Using appropriate mathematical processes and/or calculations to determine a solution

1.3 Explaining a solution in relation to the context

Outcome 2 The learner will:

2 Use reasoning skills and statistical skills linked to straightforward real-life contexts by:

2.1 Interpreting a situation involving data and identifying an appropriate strategy

2.2 Representing data appropriately

2.3 Interpreting and/or comparing data to draw conclusions

Evidence Requirements for the Unit

Assessors should use their professional judgment, subject knowledge and experience, and understanding of their learners, to determine the most appropriate ways to generate evidence and the conditions and contexts in which they are used. They should ensure that there is sufficient evidence of competence in financial, statistical and reasoning skills from the Outcomes and Assessment Standards to allow a judgment to be made that the learner has achieved the Unit. Assessors should use their professional judgment to give learners credit for an appropriate degree of accuracy. This may mean giving credit for incomplete or numerically incorrect solutions which show correct methodology, therefore demonstrating required knowledge and understanding of the financial and statistical processes involved.

Evidence may be presented for individual Outcomes or it may be gathered for the Unit as a whole through integrating assessment in one activity. If the latter approach is used, it must be clear how the evidence covers each Outcome. A calculator or equivalent technologies may be used. For this Unit, learners will be required to produce evidence as follows.

For Outcome 1 learners will be required to provide evidence of using reasoning and financial skills linked to straightforward real-life contexts by drawing on the following: determining a financial position, given budget information; investigating factors affecting income; determining the best deal, given two pieces of information; converting between currencies; investigating the impact of interest rates for savings and borrowing in a basic situation.

For Outcome 2 learners will be required to provide evidence of using reasoning and statistical skills linked to real-life contexts by drawing on the following: using statistics to investigate risk; using and presenting statistical information in diagrams; using diagrams to illustrate data; comparing data sets, using mean and range; constructing a frequency table; constructing a scattergraph; drawing a best fitting straight line on a scattergraph.

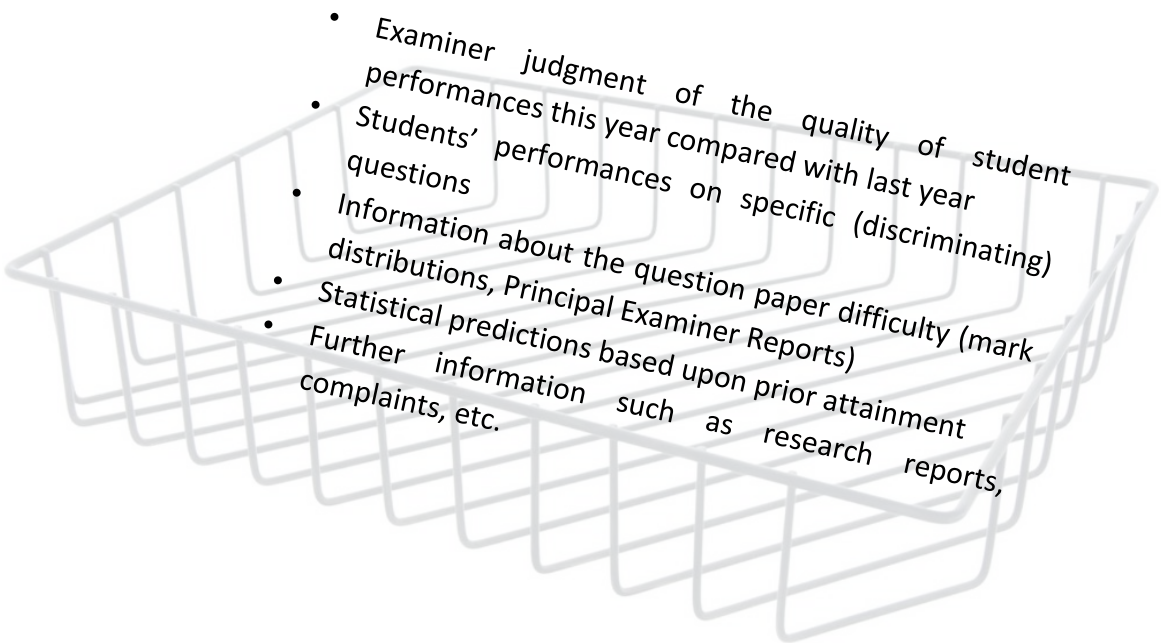
Source: SQA (2017)

3.2.1.4 Attainment-referencing

Definition Candidates receive grades that reflect their holistic attainment in the qualification at a standard which is comparable with the attainment required for that outcome in previous years' qualifications.

Method Attainment-referencing seeks to award grades on the basis of students' attainment, which can be affected by the difficulty of the question papers they have taken. As GCSEs are not pre-tested, their difficulty varies from year to year, albeit usually only by a small number of marks. Examiner judgment and statistical information are needed to ascertain the extent to which the assessment is more or less difficult, as examiner judgment alone is known to be fallible (Baird & Dhillon, 2005; Benton & Bramley, 2015; Good & Cresswell, 1988; Stringer, 2012). Equally, examiner judgment is necessary to detect any cohort-level changes in performance because statistical methods alone can simply replicate the distribution of grades from previous years. The range of evidence used to set grade boundaries is sometimes colloquially referred to as a 'basket' of information (Figure 3).

Figure 3 Example basket of evidence used in attainment-referencing



Interpretation Outcomes under attainment-referencing represent the grades that students should be awarded, given the difficulty of the question papers that they have sat in any given year. Attainment-referencing is designed to counteract changes in the difficulty of the question papers while maintaining a relationship between candidates' attainment and the grades they are awarded. Written grade descriptors may be used to explain the kinds of

knowledge and skills that candidates awarded each grade should have demonstrated in the assessment.

Origin Newton (2011) first used the term attainment-referencing. He notes (fn. ii, p. 26) that this refers to the concept previously termed weak criterion-referencing by Baird et al. (2000).

Main strengths and problems Since qualitative and quantitative information are combined under this approach to standard setting, the benefits of each source of information can accrue. However, exactly how this information is to be integrated and which source should be given most weight in final decisions about grade boundaries is also a matter for judgment. The very flexibility of attainment-referencing is a strength because it allows for judicious approaches tailored to the specifics of the context of the assessment. However, this flexibility is also a weakness since there is not a sole artefact that represents the standard. This means that the approach is open to claims of cherry picking evidence to support changes in outcomes. Statistical outcomes may appear like grade inflation, or too stringent due to genuine changes in student performances (caused, for example, by trends for changes in curriculum timetabling for a particular subject). Student performances may appear weaker or stronger due to a difficult or easy question paper. Indeed, attainment and performance may not perfectly align for many reasons: see Box 2**Error! Reference source not found.Error! Reference source not found.**.

3.3 Summary of Chapter 3

- This project focuses upon how standards are set for GCSEs in Wales.
- The method used to set standards in assessments varies and affects the inferences drawn from the results.
- This report explains how each stage in a qualification lifecycle contributes to the setting and upholding of standards in the system for GCSEs in Wales, in normal times. The report also considers what the qualification lifecycle and associated assessment processes would look like under different methods of setting standards – norm- and criterion-referencing.
- To inform the critical evaluation technique used in the project, a review of the previously published literature was undertaken and nine interviews were conducted with nine WJEC and Qualifications Wales assessment professionals.
- Three statistical methods for setting standards are explained: norm-referencing, cohort-referencing and comparable outcomes.
- **Norm-referencing** involves ranking candidates in relation to a pre-defined population and comparing individual scores to the distribution of scores in the broader population.

- **Cohort-referencing** ranks candidates based on their performance compared to that of others in the same year or series.
- **Comparable outcomes** aims to give candidates similar grade profiles to those of previous qualifications and to predict results based on prior attainment data.
- **Attainment-referencing** is a method where grades reflect students' holistic attainment in comparison to previous years' qualifications. Attainment-referencing considers the difficulty of question papers and uses both examiner judgment and statistical information. This is the method used to set standards for GCSEs in Wales in non-pandemic years.

Box 2 Attainment versus performance

The aim of attainment-referencing is to award grades on the basis of students' underlying attainment. Students' performance on assessment tasks is a key source of evidence about attainment but there are circumstances in which changes in performance (improvements or deteriorations) do not indicate changes in attainment – and vice versa. Some of the factors leading to this disconnect between performance and underlying attainment include:

2. Changes to the assessment:
    a. Changes in assessment difficulty – a more difficult assessment will elicit a weaker performance. An easier assessment will produce an improved performance.
        i. Changes in assessment difficulty may be driven by changes in question type – a question paper with more structured questions will likely elicit a stronger performance than a paper with more open-ended questions.
        ii. The removal or introduction of tiering is an extreme change to assessment difficulty. Tiered papers will likely elicit a stronger performance as very able students and the weakest students will have more opportunity to show what they know and can do.
    b. Changes in the form of the assessment – performance in coursework will likely be stronger than performance in examinations due to the opportunity for reflection and to revisit work, and because of the support available. The higher the controls around coursework, the more similar the performance will be across the two forms of assessment.
    c. The overall quality of the assessment – a well-designed assessment is more likely to elicit construct relevant performance.
    d. Aggregation effects – as the number of units of assessment and the correlations between performance on the units change, the

performancerequired to achieve a particular overall qualification grade will need to be adjusted even if attainment has not changed.

3. Changes in the classroom:
    a. Coaching – as content elements and assessment formats become familiar, teachers may teach strategies for scoring marks, so performance can improve without any increase in attainment (Newton, 2020).
    b. Adeptness – teachers may help students to become better at recognising, navigating and responding to the assessment demands (Newton, 2020). A lack of adeptness can mean that students' performances are lower than their underlying attainment.
    c. Reallocation – over time teachers become better at question spotting, redirecting their teaching to those areas that are most often sampled in the assessment (Newton, 2020). This can result in improved performance in the absence of improved attainment.

4 Embedding standards in the GCSE qualifications lifecycle

Although GCSE qualifications were first awarded in 1988, they have been regularly updated. Each time this happens, there is an opportunity to reinvigorate the standards through the design and development of the curriculum and assessment materials and the delivery processes. With assessment, seemingly small details can impact upon the qualification standards. Evaluation work is often conducted, either on major structural issues (such as the effect of modular assessment) or on more specific matters (such as the vocabulary requirements for GCSE Spanish). These evaluations then inform decisions on the subsequent design of the qualifications. We explain here how each of these stages affects standards in GCSEs currently throughout the qualification lifecycle (Figure 4).

Figure 4 GCSE qualification lifecycle

4.1 Design and develop

In these stages of the qualification lifecycle, the subject content and assessment objectives are defined and the structure of the qualifications, the assessments and assessment procedures are designed. In Wales this work is conducted across organisations, including Welsh Government, Qualification Wales and WJEC. The Welsh Government is responsible for the national curriculum framework. Qualification Wales is responsible for setting rules for the design of the qualifications.

4.1.1 GCSE approval criteria and additional rules

Qualifications Wales provide the framework and criteria against which exam boards develop specifications to be submitted for approval. The rules that apply to the GCSE suite are expressed in approval criteria and additional conditions. The rules set the grading scale and parameters for assessment, for example, stipulating that a variety of question types and tasks are used. Other design rules set the maximum number of assessment units in a GCSE and the minimum weighting for any unit. And in most modular GCSEs, the rules stipulate the number of re-sits allowed and the weighting of terminal assessment (40 per cent of the assessment must be taken in the series that the student certificates in). The terminal assessment rule is crucial in the setting of comparable standards across units and in setting an appropriate qualification level standard.

4.1.1.1 *GCSE approval criteria and additional rules and threats to standards*

The design rules are implemented to embed and protect standards. Nonetheless, the choices made have implications for standards and there is rarely clear-cut research evidence to provide guidance. For example, modular qualifications give timely feedback, allowing students to track their progress (Baird et al., 2009; Vidal Rodeiro & Nádas, 2012), are motivating (Noyes & Sealey, 2011) and may make learning easier by presenting content in a clearer way (AlphaPlus Consultancy Ltd, 2012). These factors may improve performance standards. On the other hand, there is some evidence that knowledge may not be retained after taking units of assessment (Barham, 2012) and students sitting early units are younger and so less mature than their linear counterparts (Forster, 2011). These factors may undermine performance. Indeed, the evidence surrounding the impact of modularity versus linearity on standards is mixed and it is likely that the efficacy of either approach is highly context dependent (see Baird et al. (2019) for a full review).

4.1.2 Subject-specific criteria

The approval criteria interact with subject-specific criteria, also set by Qualifications Wales. The exam boards follow these rules when creating the detail of each GCSE specification. The subject-specific rules vary across subjects and their format and detail varies according to when the qualification was reformed. Nonetheless, those for GCSE Welsh Language are used here as an example to demonstrate their role in embedding standards in the qualification (Welsh Government, 2014). GCSE Welsh Language was revised in the first phase of the last programme of reform. Each part of the rules is considered in turn:

The **rationale** for the GCSE Welsh Language says it should 'provide greater assurance of literacy' because 'the levels of literacy demonstrated by many learners are not high enough' (p. 2). The rationale goes on to say that the GCSE should 'focus primarily on the

functional aspects of language' and that students 'will develop their ability to use Welsh as active and informed citizens and be able to speak, listen, read and write fluently, appropriately, effectively and critically – for a wide range of personal, functional and social purposes' (p. 3). As such, the rationale for a GCSE plays a significant role in setting expectations about the required standards.

The **aims and learning outcomes** go on to set out what a GCSE specification will enable learners to do. For example, one of the outcomes required of students studying GCSE Welsh Language is that they 'develop their verbal reasoning and their ability to think constructively and critically in response to written and digital/dynamic texts' (p. 4). Again, the aims and learning outcomes play a key role in setting expectations about standards.

The Welsh Language **subject content** falls into three skill areas – oracy, reading and writing. In oracy, for example, learners are expected to 'speak accurately and fluently, adapting style and language to a wide range of forms, contexts, audiences and purposes' (p. 5). In reading, learners are expected to 'demonstrate verbal reasoning skills in synthesising and summarising information from a range of texts' (p. 5). As such, the subject content goes some way to lay out expected performance standards in the subject.

The **assessment objectives** are an articulation of the skills required in the context of the subject content and are given weightings to reflect their importance. In GCSE Welsh Language, oracy skills have a 30% weighting, reading skills 30% and writing skills are given a greater weighting at 40%. The assessment objectives set out an additional layer of expected standards. For example, in oracy, students are expected to 'listen and respond appropriately to other speakers' ideas, questions and perspectives, and how they construct and express meanings' (p. 6). Further requirements may be expressed, for example in writing there is a rule that half of the available credit will be awarded for spelling, punctuation and grammatical accuracy.

The **scheme of assessment** sets out high-level design principles for the qualification. These include whether the qualification is linear or modular, whether there are tiers of entry, when assessments are taken, whether/when re-sits are allowed, the number and content of assessment units, the form of the assessments for each unit (examination or coursework, for example), any constraints around the length of assessments, and any restrictions on the use of aids such as dictionaries and calculators.

The scheme of assessment may also set out details regarding how the assessments should be constructed, the type of questions used and how marks should be awarded. For example, the scheme of assessment for GCSE Welsh Language states that the oracy

assessment should comprise two tasks – an individual, researched presentation and a group discussion. The tasks should be equally weighted and half of the marks should be 'awarded for the choice of appropriate register, grammatical accuracy and range of sentence structures with the remainder for content and organisation' (p. 8). With regard to the assessment of reading, the scheme of assessment requires a mixture of short response questions (e.g. multiple-choice questions, short constructed responses, cloze, sequencing) and longer response questions (e.g. paraphrasing, context comprehension, analysis/deduction/inference). This is an important way of embedding standards as the structure of questions can impact the demand of the paper (Pollitt et al., 1998).

The scheme of assessment will have a significant effect on standards. Design features such as the number of assessment units, modularity and re-sitting not only impact the complexity of the process of setting grade boundaries but also impact the performance standards required to reflect underlying attainment. Consider the decision to create tiered assessments. Tiering affects the demand of the question papers and better allows students to demonstrate what they know, understand and can do. As such, the performance standard may increase with a tiered design even if underlying attainment is constant. Similarly, assessment by coursework may well generate work at a higher performance standard than that produced under exam conditions, even when the level of attainment is the same.

Grade descriptors exist only for those GCSEs included in the first phase of reform. Their aim was to help teachers by providing an indication of the likely level of performance at key grades (A, C, F). The descriptors reflected the fact that the compensatory nature of the assessment means that shortcomings in some aspects of students' performance may be balanced by better performances in others. Compensation has advantages in terms of fairness to students and flexibility (Cadwallader, 2014). Students who slip up against simple criteria are not overly penalised. However, compensation undermines the extent to which grade descriptors can be used to set standards. Successes and failures can be combined in a multitude of ways to achieve a specific mark, meaning that two candidates could achieve the same grade but exhibit very different patterns of performance. As a result, it is difficult to describe the 'typical' performance associated with a given grade; there are simply too many possible definitions, each of which is based on the exact way in which a candidate reaches a given range of marks (Baird & Scharaschkin, 2002; Cresswell, 1987). These difficulties were part of the rationale for moving away from grade descriptors.

4.1.2.1 *Subject-specific criteria and threats to standards*

These rules do not state whether aims and learning outcomes describe a minimum standard, a standard expected of a typical student or that expected of the most able.

That is not their purpose. To take assessment objectives as an example, these could, with no or minor adaptation, be used to describe the skills required in a qualification of a higher or lower level. To 'listen and respond appropriately to other speakers' ideas, questions and perspectives, and how they construct and express meanings' requires a quite different standard of performance at grade A compared to grade G. As such, the subject rules are open to considerable interpretation. It is through the creation of specifications and sample assessment materials that the standard becomes more concrete.

The way in which the information expressed in subject criteria can lead to different interpretations of the expected standard was exemplified in problems encountered during the accreditation of England's GCSE Mathematics specifications in 2015. There were very significant differences between exam boards in the difficulty of the sample assessment materials. Indeed, most of the papers were highly likely to have proved too difficult for all but the very best students, whereas others were likely to have been slightly too undemanding (Ofqual, 2015a).

4.1.3 Approval

To be funded for use in schools and colleges in Wales, a qualification needs to be regulated by Qualifications Wales. Only recognised exam boards may offer GCSEs. To be recognised, the exam board must demonstrate compliance with a suite of regulations set by Qualification Wales – the Standard Conditions of Recognition (Qualifications Wales, 2021b). They are general rules that, among other things, underpin the validity of qualifications. For example, there are rules intended to ensure that the demand of question papers is consistent over exam series; the language of assessments enable students to demonstrate their attainment; and the confidentiality of assessment materials is maintained.

Recognised exam boards can submit GCSEs to Qualifications Wales for approval. GCSEs must go through this process of approval before they can be made available for teaching. To be approved, a specification must meet criteria expressed in a series of regulatory documents: Approval Criteria for GCSE Qualifications; Subject Level Criteria; Additional Standard Conditions of Recognition for GCSE/GCE Qualifications; and the Standard Conditions of Recognition. Qualification Wales will only approve a qualification that meets all of the requirements set out in these documents.

In the process of approval, the exam board submits the qualification materials (specification, rationale and sample assessment materials) to Qualification Wales. Trained subject experts review the materials against the approval criteria and write a feedback report. Any areas of non-compliance are fed back to the exam board. If needed, the exam board resubmits the qualification with a response to the feedback.

The submitted materials are reviewed again. This process is repeated until Qualifications Wales is satisfied and approves the qualification. The process of approval is designed to ensure that GCSEs meet the requirements not only at the point of approval but across the lifespan of the qualification. As such, the process plays a pivotal role in embedding and protecting standards.

4.1.3.1 Rationale and Sample Assessment Materials

The rationale submitted by the exam board is an explanatory document explaining the rationale for the key design features. Its aim is to support the qualification review and approval process. Where features have not been prescribed, it must include an explanation for the following:

- the qualification structure, explaining the reasons for the way in which the subject content has been organised;
- the qualification content, for example an account of how and why texts or authors have been selected;
- how the requirement for including a Welsh perspective has been met and how the approach taken is appropriate to the subject;
- the assessment structure, including the number, weighting, mark allocation and duration of assessment units;
- how the spread of assessment objectives across and within the assessment questions was decided;
- the design of sample assessment materials, including the type and range of assessment tasks in each unit and their relationship to the assessment objectives;
- the design and application of the mark scheme.

Sample assessment materials are submitted for coursework tasks and examinations, including mark schemes. They play an important role in the approval process as they exemplify the assessments the exam board plan to produce, they provide a broad template for future assessments, and, in the absence of past papers, schools and colleges use them to guide their teaching. After all, the better prepared teachers are to teach the new specifications, the better the standard of student performance.

4.1.3.2 Approval and threats to standards

The approval process is intended to ensure that qualifications comply with regulations and so are likely to have appropriate content, assessment and grading standards. While an exam board submitting a qualification must demonstrate that the qualification will be compliant over time, continual monitoring during delivery is also used to provide longer-term assurance. Indeed, approval is given in the absence of teachers teaching the subject content and students sitting the assessments, so it can only provide important but limited assurance that standards will be appropriate.

4.2 Delivery phase

In the delivery phase a measurement result in the form of a mark and a grade is delivered for each candidate both for individual assessment units and for the qualification overall.

4.2.1 Setting the assessment and associated mark scheme

Assessment may be by examination or by coursework. In the development phase a broad template for the design of the assessments is created in the form of sample assessment materials. In the delivery phase the sample assessment materials are referred to in the creation of 'live' assessments. Where an assessment is available in more than one language (Welsh and English, for example) the assessments in different languages must be of comparable demand.

4.2.1.1 Examinations

The aim is to produce exam papers of the same demand as previous years, covering subject content and assessment objectives in line with the regulations. Regulations may specify the style of the question paper, including the types of items (multiple choice, short response, essays) that will be used. Where there is optionality, the level of demand must be comparable across choices. A tracking system is used in which assessment objective coverage is mapped across questions to ensure their appropriate weighting. Subject content coverage is also tracked over time and the proportion of different question types is monitored.

The mark scheme is produced at the same time as the assessment, as the two must work hand in hand. The assessment must elicit proper evidence of attainment in the subject, and the mark scheme must enable the evidence to be evaluated properly (Ahmed & Pollitt, 2011).

A number of experts are involved in the process. The Principal Examiner writes the questions and reviews the paper at certain points during its production. The Reviser checks that the questions are appropriate and error free. The paper is then reviewed by a committee of senior examiners and subject experts to ensure the questions are clear and that the assessment is valid. At least one member of this committee is a Welsh speaker.

The Scrutineer then sits the question paper as if they were a learner. They check it can be completed in the time allowed, that there are no errors and that the level of demand is correct. The Translator translates the English version of the paper into Welsh. A subject specialist seeks to ensure that the Welsh translation matches the English version and that the terminology matches the specification.

4.2.1.2 Coursework

Coursework tasks (such as presentations, essays and portfolios) are designed to assess students' performances against the assessment criteria set out in the specification for the subject. Although all students are assessed using the same criteria, subject to approval from the exam board, there is sometimes scope for the schools to set the topics that the tasks are based on. Where students can choose between tasks, the level of demand must be comparable. Coursework tasks typically do not change from year to year.

4.2.1.3 Assessment setting and threats to standards

4.2.1.3.1 Demand of assessments

The aim is to ensure that the demand of assessments is consistent over time and across any options. This can be difficult to achieve without pre-testing of questions. Research has shown that examiners involved in question setting are unable to accurately predict how difficult students will find them (El Masri et al., 2017). Even small changes, such as to the structure of questions, can impact the demand of the paper. For example, questions that are broken down into parts tend to be easier for students to score marks on than unstructured questions (Pollitt et al., 1998). This is why grade boundaries need to change over examination series. Further, tightly managing question paper demand over time, for example by replicating question structure and content coverage, needs to be balanced against the risk that the question papers become overly predictable.

4.2.1.3.2 Mark scheme construction

Mark schemes must be constructed to support reliable marking and to reward features in students' responses that reflect the intended attainment construct. The function of a mark scheme is to help markers to distinguish between better and poorer responses to questions, and to determine the boundaries along the continuum of performance where the number of marks awarded should change (Ahmed & Pollitt, 2011). To the extent that marks are awarded for things that are not evidence of learning, or not awarded for what is evidence of learning, standards will be undermined (Pollitt et al., 2008). Insufficient attention to mark scheme construction is a risk and this is why it is important that the assessment and mark scheme are jointly constructed.

4.2.1.3.3 Intended weighting of assessment objectives

The intended weightings of assessment objectives reflect the relative importance placed on them by those setting the subject content and examining the subject. They are part of operationalising content standards. The intended weightings are normally reflected in the number of marks assigned to each assessment objective. It is important that these weightings are reflected in the design of assessments. If the achieved

weightings are very different from those intended, the most successful students may not be those who achieve the intended balance of knowledge and skills but rather those whose performances lean towards certain types of knowledge or skills (Stringer, 2014). This would constitute a threat to standards. For example, if items related to one assessment objective prove too easy for students, it is unlikely that the assessment objective will achieve its intended weight.

4.2.1.3.4 Familiarity

Newton (2020) lists the reasons why we might expect performance and/or attainment to rise over time over the lifetime of a specification. As new content elements become familiar, teachers become better at teaching them, resources improve, and consequently learners come to learn them better. He calls this realignment. Realignment can result in improvements in both performance and underlying attainment. As such it is not a threat to standards.

As new assessments become familiar, teachers will help learners to become better at recognising, navigating and responding to the new demands, so students become more effective at demonstrating their actual levels of attainment. Newton calls this adeptness. A lack of adeptness can threaten standards. Adeptness is distinct from realignment because increasing adeptness will result in students becoming better at demonstrating their levels of attainment, even if their levels of attainment are constant over time.

Newton notes that increasing adeptness could be less innocuous if one follows Shepard's (1997) line of thinking. Shepard argued that as new task formats become increasingly familiar to students, they become more likely to be able to score marks with only a weak grasp of the subject material. Again, this might occur even if level of attainment remained constant over time and as such it could threaten standards.

Newton notes two other effects which can undermine standards. Coaching can occur. As new content elements and assessment formats become familiar, teachers begin to teach strategies for scoring marks and so performance can improve without any increase in attainment. And reallocation can also occur. As new content elements and assessment formats become familiar, teachers become better at question spotting, reallocating their instructional resources towards those areas that are most often sampled. This can result in improved performance in the absence of improved attainment. Reallocation is related to the predictability of the assessment.

4.2.1.3.5 Predictability

Exams should be sufficiently predictable to enable students and teachers to have enough of an expectation of the nature of the demands and coverage to manage

preparation ('test wiseness') and test anxiety. However, if there is too much predictability, as outlined above, expectations can lead to narrowing of preparation and even rote learning of responses (Baird et al., 2016; Holmes et al., 2020). As such, undue predictability may dilute standards.

4.3 Exam papers are delivered to centres

During delivery and storage, it is essential that assessment materials are kept confidential. Question papers must be kept in their sealed packets in a secure storage facility.

4.3.1 Exam paper delivery and storage and threats to standards

Security breaches risk malpractice or the suspicion of malpractice. On some occasions it is necessary to replace assessments at short notice or to provide a replacement paper to some of the student cohort (if, for example, a paper is intentionally or unintentionally made available before the exam). This may challenge standard setting, for example if the papers are of different levels of demand.

4.4 Coursework is conducted

To support standards, there are rules concerning the way in which coursework is conducted. These rules vary according to the subject and coursework task. Some level of control is usually needed to ensure that the work is authentic and that the demands placed upon students are consistent. But, in the interests of validity, some coursework is completed by students with very few limitations on the task set or the environment in which the task is taken. It may be appropriate, though, for coursework in other subjects to be completed in highly controlled conditions at a set time and place. Most coursework is conducted under rules somewhere within this spectrum of control. For example, in GCSE English Language, candidates are given a set amount of time to prepare for a task. During this time, they work under limited supervision and teachers are allowed to offer guidance and advice to students about undertaking the task.

4.4.1 *Conducting coursework and threats to standards*

4.4.1.1 Consistency of support and demand

There is a risk that schools and colleges come to different interpretations of the amount of assistance teachers can give to students and the different degrees to which teaching focuses on specific assessments (Opposs, 2016). This can create inconsistent standards between schools and colleges which the marker moderation process cannot easily detect.

4.4.1.2 Familiarity

As teachers and students become increasingly familiar with coursework tasks and as more examples of responses are available, the demand of the tasks may diminish. The effects of familiarity outlined by Newton (2020) (see above) are especially prominent given that the tasks often stay the same.

4.4.1.3 Malpractice and maladministration

Where there is room for interpretation of the rules, teachers may (consciously or unconsciously) give too much support to their students (Meadows & Black, 2018; Opposs, 2016). Opposs also points to the risks of plagiarism and excessive input from parents as threats to coursework standards.

Essay mills can also undermine standards in the coursework elements of qualifications. In 2018, a study suggested that, globally, up to 15.7% of college students have used such a service, with the rate rising rapidly over the past few decades (Newton, 2018). However, this form of cheating may become obsolete with recent developments in artificial intelligence in which 'chatbots' can use neural networks to write sophisticated responses which are difficult to spot (King & chatGPT, 2023).

4.5 Exams are conducted

Students must sit the exam at the designated time in an exam room with the right conditions – appropriate temperature, lighting, ventilation and noise levels. The seating must prevent students from overlooking (intentionally or otherwise) the work of others. Invigilators are responsible for conducting the exams. They play a key role in upholding the integrity of the exam process and must not be involved in the teaching of the examined subject. They must keep exam materials secure and prevent candidate malpractice. Exam timetable clashes need to be managed in such a way as to prevent malpractice, even including overnight supervision of students if necessary.

4.5.1 *Reasonable adjustments for disabled learners*

Reasonable adjustments must be made so that disabled students can demonstrate their knowledge, skills and understanding. A reasonable adjustment can be unique to an individual student, although certain types of reasonable adjustment are more commonly made. Some of the most frequently made reasonable adjustments include extra time, a scribe to write down a student's dictated answers, access to assistive technologies and provision of the exam paper in an enlarged font. However, there are limits to the range of allowed adjustments, as they must not undermine the validity of the assessment. For example, the use of a human reader in the assessment of reading ability would not be permitted.

4.5.2 Special consideration

If a student temporarily experiences an illness, injury or other event outside of their control immediately before or during an exam, they may be given special consideration. For example, the conditions under which the exam is taken may be adjusted, a small number of extra marks may be awarded, or the qualification may be awarded even though the student was absent for one of the exams.

4.5.3 Conducting the exam and threats to standards

There are a number of threats to standards at this stage of the assessment process.

4.5.3.1 Malpractice and maladministration

Instances of detected malpractice represent a very low proportion of examinations taken. In 2022, approximately one penalty was issued for every 4,200 entries (Qualifications Wales, 2022a). Nonetheless, student and teacher malpractice can undermine standards, and the range of ways to cheat is continually evolving (Independent Commission on Examination Malpractice, 2019).

4.5.3.2 Reasonable adjustments and special consideration

In a system of this scale, it can be challenging to ensure that all students who need reasonable adjustments receive the correct adjustment (Hipkiss et al., 2021). For example, there has been debate around the numbers of students receiving extra time and, perhaps more importantly, whether the right students are receiving the right amount of extra time (Independent Commission on Examination Malpractice, 2019). Indeed, manageability constraints mean that students with different disabilities are given set amounts of extra time and the evidence base for these amounts is weak (McGhee & Masterson, 2022). All of these concerns also apply to special consideration and to the extent to which they manifest, they pose a threat to standards.

4.6 Marking exam papers

Exam scripts are usually marked onscreen by examiners. Examiners are usually teachers, recruited based on their subject expertise. Sometimes they mark batches of items, and at other times they are given batches of whole scripts to mark. Before they begin marking, the Principal Examiner for the unit holds a standardisation meeting. Even at this stage amendments can still be made to the mark scheme where necessary so that worthy learner responses are credited. At the meeting markers are trained to use the mark scheme. They mark a common set of scripts and review their marks to ensure they are marking consistently. Before they are allowed to begin marking, they are required to mark a set of 'qualification' items or scripts sufficiently accurately.

During marking, each examiner's marking is checked to ensure that it is consistent and to the required standard. Where marking is done onscreen, 'seeding' or double marking is used. In the former system, seeded items are included through the marking. Seeds are student responses that senior examiners have reviewed and for which they have agreed a mark. Examiners don't know which items are seeds.

Double marking may be used for longer response items. A sample of the batch of responses is marked by another examiner. If the marks of the two examiners are not within an agreed tolerance, a senior examiner adjudicates.

In both quality assurance systems, examiners will be stopped from marking if they do not mark to the agreed standard to receive training from a more senior examiner. If they continue to misapply the mark scheme, they may be stopped from marking altogether.

Where marking is done on paper, examiners send samples of their marking to a more senior examiner for checking. Again, if the examiner is not marking to the required standard after additional training/feedback has been provided, they are not allowed to continue and their allocation is given to another examiner.

4.7 Marking coursework

Teachers in a school or college mark their students work against the assessment criteria provided by the exam board. When more than one teacher is involved in marking an assessment, the markers will seek to standardise their marking before they begin. Once marking is complete, their marks are sent to the exam board and a sample of marked work is subject to a moderation process to check that the marks have been awarded in line with the agreed standard.

Moderators undergo standardisation to ensure that they have a shared understanding of the marking criteria. This typically involves individual and group scrutiny of a number of non-live responses that have been pre-selected by the principal moderator.

If the original marks are consistent with those of the moderator (within a specified tolerance) then the original marks are accepted. If the marking is outside the tolerance, the moderator reviews a further sample and the marks are analysed to determine whether the marks from the school or college need to be adjusted. The purpose of moderation, therefore, is not to re-mark individual responses but to align standards across schools and colleges. Moderators' work is checked at regular intervals by senior moderators to ensure that their judgments are consistent and in line with the agreed standard.

4.7.1 Marking and threats to standards

The mark scheme embodies the knowledge, understanding and skills that students must learn – the demands of the unit. Ensuring that the mark scheme is applied correctly – that student responses are rewarded correctly – is part of embedding content standards. If inappropriate aspects of students' work are rewarded then the rank order is disrupted, undermining standards.

If some teachers are too generous or severe in marking coursework and the moderator becomes anchored to the teacher's marks and so fails to identify that an inappropriate standard has been applied, standards may be undermined (Cuff, 2017). The system is therefore reliant on successful quality assurance of marking, be it of exam papers or coursework.

4.8 Grading

Once marking is complete, it is possible to determine grades. An awarding meeting is convened to recommend grade boundary marks for key grade boundaries in each subject. These are A, C and F in non-tiered GCSEs and A and E at A-level. The grade boundaries for intermediate grades are calculated arithmetically. Awarding committees are chaired by a senior examiner who has responsibility for standards in the subject. The committees also include Principal Examiners (responsible for examined units), principal moderators (responsible for coursework units) and exam board technical experts. The meeting may be held face to face or online.

The principle behind the awarding process is to maintain, year on year, the level of performance at a grade boundary mark. While it is intended to design exam papers at the same level of demand as in previous years, there is no pre-testing of exam questions and so in practice this is very difficult. This means that grade boundaries have to be adjusted according to the demand of the exam. A variety of sources of evidence are used when setting boundaries – both statistical and judgmental. This approach would be described in the academic literature as attainment-referencing.

At GCSE the main statistical evidence[5] takes the form of the outcomes of centres that are common to the reference year and the current year in a subject (see Box 3). The

[5] At A-level the main statistical evidence is predictions based on prior attainment at the cohort level. These predictions map the relationship between prior attainment (mean GCSE score) and A-level outcomes for the cohort of students taking each subject in a reference year. This relationship is used to

assumption is that the outcomes for these common centres will be stable from one year to the next. On this basis, results can be predicted for the common centres. The predictions guide the setting of boundaries, which are then applied to all students. In the interests of accuracy, there are sometimes adjustments to the way in which the common centres are selected. For example, only common centres with stable entry sizes might be selected (Pinot de Moria, 2020a). Where a GCSE is tiered (e.g. mathematics) efforts are made to align the performance standard over time but also across tiers (e.g. grade C on the Higher and Intermediate tiers).

Box 3 Comparable outcomes – a worked example using common centres

There are five centres with entries for the GCSE in the reference and current year. If the number of candidates entered by each of the centres for the current year was the same as for the reference year, the assumption underpinning the common centres prediction would mean that the reference year overall grade distribution would become the prediction for the current year. However, this is unlikely ever to be the case. In this example the prediction for the current year is lower than the actual grade outcome in the reference year. That is because the lower-achieving centres (1 and 2) have a greater number of candidates entered for the qualification in the current year and the higher-achieving centres (4 and 5) have a lesser entry.

Reference Year Outcomes (e.g. 2018)

Centre	Entry	A*	A	B	C	D	E	F	G	U
1	13	7.7	23.1	46.2	69.2	76.9	92.3	92.3	100.0	100.0
2	21	4.8	23.8	47.6	66.7	85.7	95.2	95.2	100.0	100.0
3	25	8.0	24.0	44.0	68.0	84.0	96.0	96.0	100.0	100.0
4	29	10.3	27.6	48.3	72.4	82.8	93.1	96.6	100.0	100.0
5	33	12.1	27.3	51.5	72.7	84.8	93.9	97.0	100.0	100.0
Overall	**121**	**9.1**	**25.6**	**47.9**	**70.2**	**83.5**	**94.2**	**95.9**	**100.0**	**100.0**

predict the outcomes for the current cohort of students based on their prior attainment. If the prior attainment of the cohort remains similar, the outcomes are expected to be similar.

The predictions for each A-level cohort are usually created for 18-year-old students and are therefore based on the GCSE results that the students gained two years earlier. The predictions guide the setting of boundaries, which are then applied to all students.

Current Year Prediction (e.g. 2019)

Centre	Entry	A*	A	B	C	D	E	F	G	U
1	33	7.7	23.1	46.2	69.2	76.9	92.3	92.3	100.0	100.0
2	29	4.8	23.8	47.6	66.7	85.7	95.2	95.2	100.0	100.0
3	25	8.0	24.0	44.0	68.0	84.0	96.0	96.0	100.0	100.0
4	20	10.3	27.6	48.3	72.4	82.8	93.1	96.6	100.0	100.0
5	12	12.1	27.3	51.5	72.7	84.8	93.9	97.0	100.0	100.0
Prediction	**119**	**7.9**	**24.7**	**47.0**	**69.2**	**82.4**	**94.1**	**95.0**	**100.0**	**100.0**

As an example, the calculation of the grade A* prediction is set out below:

$$Grade\ A^*\ Prediction = \frac{\sum_i \left(Entry_{i,year(n)}\ x\ Grade\ A^*_{i,year(n-1)}\right)}{\sum_i \left(Entry_{i,year(n)}\right)}$$

$$= \frac{[(33\ x\ 7.7) + (29\ x\ 4.8) + (25\ x\ 8.0) + (20\ x\ 10.3) + (12\ x\ 12.1)]}{119}$$

$$= 7.9$$

$$Where: year(n) = current\ year$$

$$year(n-1) = reference\ year$$

$$i = common\ centre$$

Adapted from Pinot de Moria (2020a)

In addition to statistical information, judgmental evidence is used to set boundaries. The awarding committee receives reports from the principal examiners and moderators and descriptions of the expected level of performance at each key grade. They are presented with exam scripts in a range of marks (typically three to five) as guided by the statistical evidence and must independently decide whether each exam script is worthy of the grade being considered. They do so with reference to archive scripts on the grade boundary marks from previous years and statistical evidence showing the performance of individual questions on the exam paper. The judgments for each script are recorded as ticks, question marks or crosses on a 'tick chart' (see for example, Taylor & Opposs, 2018). Based on the balance of ticks and crosses, the chair of examiners specifies a 'zone of uncertainty' – that is, the range of marks within which the grade boundary is likely to lie. The chair of the awarding committee weighs the statistical and judgmental evidence, takes advice from the committee and the exam board technical experts, and recommends the final grade boundary. Student performance on each unit is aggregated to give the final qualification grade. The precise method of aggregation varies according to the structure of the qualification (number of units, weighting of units, tiering and so on).

When setting grade boundaries, the chair of examiners considers the overall qualification level outcomes. Afterall, it is the overall qualification grade that has

currency. The modular nature of Wales's GCSEs and A-levels requires a mechanism for combining marks that students have achieved in different examination series. This is achieved by the uniform mark scale (UMS) that standardises students' marks before combining them.

The grade boundaries recommended by the chair of examiners are then sent to the responsible officer of the exam board, who has overall responsibility for the decisions. The responsible officer reviews the outcomes, considering any issues raised by the awarding committee and taking account of external information such as results in other subjects. Unusually, the grade boundaries may be moved at this point but with the Chair of Examiners' agreement and normally within the zone of uncertainty on the tick chart.

Qualifications Wales sets out the requirements for the awarding process in Wales and monitors it to check it is being undertaken appropriately. Before any grades are announced, awarding bodies report their results to Qualification Wales. If the overall results are different to what was expected, the awarding body may be asked to explain why. If Qualification Wales is not satisfied, it may ask the awarding body to look again at its processes or to conduct additional analysis. An example of an award where examiners judged that the performance of candidates required different boundaries to those expected based on statistics was GCSE Applied Science in 2019, where grade C outcomes were 5.1% higher than statistically predicted. The evidence submitted to Qualification Wales demonstrated that standards had nonetheless been maintained.

4.8.1 Grading and threats to standards

As with other elements of the assessment process there are threats to standards.

4.8.1.1 Balancing statistical and judgmental evidence

The process of maintaining grading standards relies on a combination of statistical and judgmental evidence. To the extent to which statistical evidence is over-relied upon, changes in performance will not be recognised. Equally, over-reliance on judgmental evidence may undermine the maintenance of standards, as examiners have been shown to give students the 'benefit of the doubt' (Stringer, 2012). Statistical evidence can be weak when the entry to a unit and/or qualification is low or is changing over time. Equally, judgmental evidence can be undermined by changing aggregation effects which can occur in modular qualifications.

4.8.1.2 Modularity

There are unitised and linear GCSEs on offer in Wales. Grading modular qualifications can be more complex than grading linear qualifications (Baird et al., 2019). In modular qualifications students may sit units at different times through the course. Certificating students are therefore likely to have sat different question papers that require different

grade boundaries. It is important that unit standards are comparable so that students are not advantaged or disadvantaged depending upon when they take each assessment. This comparability of standards is difficult to achieve. One reason for this is that generating statistical predictions to guide grade boundary setting is complex in modular qualifications.

4.8.1.2.1 Student characteristics

The first issue relates to the likely differences in the characteristics of students who are sitting an early unit and students certificating at the end of the course. Differences in performance are likely for a number of reasons. First, students entering at different time points will have had different amounts of teaching and exposure to the subject content. Second, students are likely to mature during the two-year course and so perform differently depending upon when they take the assessment (Clarke, 1996; Taverner & Wright, 1997; Vidal Rodeiro & Nádas, 2012). Third, students sitting the unit at different times are likely to have different levels of motivation. Students taking an early unit may take the attempt less seriously if they know that there is an opportunity to re-sit (Heinrich & Stringer, 2012). Finally, some schools might use early units as mock exams to assess how their students are progressing. The extent to which these factors might influence performance is difficult to quantify, meaning that generating appropriate statistical predictions at the unit level is challenging.

Examiners can find it difficult to identify the full extent of these differences in performance (Newton et al., 2007). They have to form judgments about the difficulty of the examination papers versus the preparation of candidates without knowing how these candidates would perform on subsequent assessments and how this would affect their overall grades (Baird et al., 2019).

4.8.1.2.2 Aggregation effects

Grade boundaries must be set for each unit of the qualification, each time a unit is available, such that overall qualification standards are maintained when unit results are aggregated. While much is known about the factors affecting the aggregation process, such as regression to the mean,[6] it is not possible to entirely predict the effects of grading each unit upon the overall qualification standard. Only when the results come together and the full data is known are the implications of the grading of each unit fully

[6] Regression to the mean is an effect that occurs when student performance on units is not perfectly correlated. The weaker the correlation, the greater the effect. At the top grades, outcomes on units will be higher than outcomes at qualification level. At the bottom grade, outcomes on units will be lower than outcomes at qualification level. This is because a student who does extremely well on one unit is most likely to perform somewhat less well on the other units. And a student who does very badly on one unit will most likely perform better on the other units. How well units correlate, and so how large the regression effect is, depends on a multitude of factors and so is difficult to predict.

available. It is possible to model the likely effects of aggregation, then adjust the unit standards accordingly but modelling is somewhat imprecise.

4.8.1.2.3 Re-sitting effects in standard setting

Currently, students are allowed to re-sit each unit in a modular GCSE once. Students typically improve their performance when they re-sit units (Vidal Rodeiro & Nádas, 2012). The facility to re-sit means that even if standards have been set appropriately on each unit initially, outcomes are likely to shift as some students choose to re-sit the assessments and perform differently. This can only be in an upwards direction, since students receive their best mark as their final mark. However, candidates can only re-take each module once before being required to take all modules again (known as a 'fresh start'; WJEC, 2022).

Quantifying the effects of re-sitting and taking this into account when setting grade boundaries is difficult. It is not known how many students will choose to re-sit, which students will re-sit (see Vidal Rodeiro & Nádas, 2012), when they will re-sit or how they will perform.

4.8.1.2.4 Low weighting of final assessment – banked marks

A further issue in setting grade boundaries in modular qualifications relates to 'banked' marks. As grade boundaries are set each time a unit is available, results are provided to schools and students at the same time. These standardised marks essentially become 'banked' in the sense that they cannot change (unless students re-sit, when their mark could increase). The 'terminal rule' helps to mitigate the risk of 'banked' marks compromising standards.

In general, the terminal rule requires students to sit at least 40 per cent of the overall assessment in the series that they certificate in. However, the rule does not dictate which 40 per cent of the assessment this should be. Therefore, depending on the structure of the qualifications, it might be possible for students to use different modules as their terminal assessment. This can introduce complexities for maintaining standards, since changing the grade boundaries on different units will impact differently on students, depending on which unit students are using as their terminal assessment – this will influence the final rank ordering of students.

The terminal rule can be particularly problematic when it is possible to use different forms of assessment – in particular, internal assessment – as the terminal module. The practice of setting grade boundaries on each module at the time that students sit the assessment means that schools and students are fully aware of how students are performing, even to the extent that, for some qualifications, they are able to calculate the number of marks that a student needs to achieve in their final module to achieve a

certain grade overall. This can be problematic when the final assessment could be coursework or controlled assessment (Baird et al., 2019). This design is therefore usually avoided.

4.8.1.3 Alternative syllabuses

Standard setting is inherently more complex when both modular and linear versions of a qualification are available. In Wales there is not a choice of modular and linear specification within a subject area. Where this has happened, for example in the past, it was possible for students to sit some of the units of the modular qualification, then, depending on their performance, choose not to certificate in this qualification, instead favouring the linear qualification where all the assessment is at the end of the course. This approach can be beneficial to students since it essentially 'wipes the slate clean' and allows students to sit the whole assessment without carrying forward any of their marks from the early units. It is therefore likely to be favoured by students who under-performed on the early units and were not on course to achieve their target grade (Taylor, 2016).

A consequence of this, however, is that the subset of students who certificate in the modular qualification are unlikely to be representative of those who sat the early units). This means that while appropriate standards may have been set on the early units for the whole cohort of students entering them, these students are unlikely to be representative of those that choose to certificate. As such, inappropriate unit standards may have been set when only the students that go on to certificate are considered. This can result in the final (or terminal) assessment standards being distorted in an attempt to control overall qualification standards.

4.8.1.4 Intended versus achieved weight of units

Standards can also be undermined if the achieved weighting of units significantly differs from what was intended. This means that attainment in the specific knowledge, skills and understanding assessed by a unit may not be reflected in the overall grade achieved by the student. For example, when large numbers of students achieve high marks on coursework units and so the distributions of coursework marks are negatively skewed, perversely performance on the coursework may have much lower than intended impact on the qualification grade received.

4.9 Issue of results

4.9.1 Post-results reviews and appeals

If a teacher or student believes that an error has been made in the marking or moderation, the school or college can request access to the exam script in question. They can ask for a clerical re-check that all marks have been included and added up

correctly. They can also request a review of the marking or moderation. The review is to check that the work was marked accurately and in line with the mark scheme. Marks are changed if an error has occurred. Errors include administrative mistakes, such as not adding up question totals correctly, or mis-entering a mark, failures to apply the mark scheme correctly and unreasonable academic judgments. Marks should not be changed if they reflect a reasonable academic judgment about the quality of the response. Marks could go up or down or stay the same following a review.

If, following a review of marking, the teacher or student believes there is still an error, the school or college can submit an appeal. The exam board will then look at the work again and decide if the mark or grade needs correcting. Following an appeal, the student's grade could go up, go down, or stay the same.

If the teacher or student still believes there is an uncorrected error, they can appeal to the regulator for a review of whether the exam board complied with regulations and their own policies and procedures in handling the appeal.

4.9.2 Post-results reviews and appeals and threats to standards

During the marking review process, it is essential that the mark scheme is interpreted correctly and in the same way as during live marking. There are aspects of the review process that make this difficult to achieve. There is a period between live marking and reviews during which cognisance of the marking standard may wane. Reacquainting reviewers with the marking standard is important. The task of reviewing is also cognitively more complex than that of marking (Howard & Black, 2017). Reviewers may become anchored to and overly influenced by the original marks awarded. This may make them reluctant to suggest large mark changes. However, their natural desire to give the student the benefit of the doubt may lead to small mark changes where, in fact, errors have not occurred (Ofqual, 2015b). This risks a different standard being applied to the work that is subject to review.

4.10 Review phase

In the review phase, evaluations are conducted which may be routine or conducted in response to specific issues. The latter may be instigated and/or carried out by the exam board or the regulator. As standards are embedded at many points across the lifecycle, relevant explorations take many forms. They often comprise what Newton (2019) refers to as micro-validation – investigations into the degree to which an assessment procedure has validity built into it by design. But they can also involve macro-validation which focuses directly upon the overarching measurement claim – does the qualification measure what it is intended to measure?

Examples of routine monitoring include analyses of how question papers have functioned (mean mark, spread of marks, spread of grade boundaries and so on), and how individual questions have functioned (facility and discrimination). The findings from these analyses are fed into the next cycle of question paper production. Exam boards may also consider data from a range of other sources. For example, from the quality assurance of marking, mark changes resulting from the marking review process, question paper errors and detected instances of maladministration and malpractice.

Exam boards will also consider feedback from schools and colleges, for example if it would be helpful to clarify an aspect of the assessment criteria for a coursework task. Exam boards or the regulator may conduct annual surveys of public, teacher and employer confidence in qualifications (see, for example, Beaufort Research (2022)).

Bespoke evaluations may be conducted in response to issues which arise during an assessment series. For example, comparability studies may be conducted in which more intensive scrutiny of performance standards and how they have changed over time is undertaken. Evaluations may also be conducted in response to public concerns that are more enduring, for example investigations into inter-subject comparability of standards.

In light of findings from these evaluations, qualifications and their assessments may be modified within the regulations. However, more substantial changes may require revision to regulations which is possible between major rounds of reform. In either case, schools and colleges need to be given sufficient time to prepare for any changes before they are reflected in the assessments.

4.11 Summary of Chapter 4

- Qualifications Wales sets rules for the design of qualifications, including GCSEs. The design rules aim to embed and protect standards and have implications for performance standards.
- Standards are embedded in qualifications throughout their lifecycle: through the design, development, delivery and review phases.
- This chapter of the report has outlined how threats to standards arise and are managed in the current procedures.
- The standard setting process is key to ensuring performance standards.
- Statistical and judgmental evidence is used to set grade boundaries for GCSEs in Wales.
- Statistical evidence for GCSE standards sometimes takes the form of common centres with stable entry policies.
- Judgmental evidence includes reports from examiners and exam scripts.

- Threats to standards in the standard setting phase include balancing statistical and judgmental evidence, modularity, student characteristics, aggregation effects, re-sitting effects, low weighting of the final assessment, alternative syllabuses, and intended versus achieved weight of units.

5 The effects of norm-referencing GCSE qualification standards on assessment processes

Here we consider how a norm-referencing approach might operate were it to be introduced for setting and maintaining standards for GCSEs in Wales. We elucidate the impact that this fundamental change would have at each stage of the qualification lifecycle, unpacking the potential risks and benefits that may be introduced. Norm-referencing can be operationalised in different ways. Many decisions would be required to define how it would work, and each of these decisions would have implications for the development and delivery of qualifications. While there are many norm-referenced tests in use, we have been unable to identify an example of norm-referencing being used to set standards in any qualification around the world.

Let us revisit our working definition for norm-referencing:

Definition Candidates receive grades that tells us where they rank in relation to the population of students who could have taken the qualification in any year.

Method To construct norms for an assessment, carefully conducted studies are carried out on representative samples of the population of interest. For example, intelligence tests can be normed for adult populations by testing a representative sample of the population with the tests and deriving what proportion score at each level. The scores of future test-takers are then compared against this established population level distribution.

Norm-referencing is often conflated with cohort-referencing (Newton, 2022). Cohort-referencing is a relational approach whereby each year a pre-defined proportion of the cohort is awarded each grade (for example, the highest attaining 8% receive an A* grade, the next highest attaining 12% receive a grade A, etc.)**.** Norm-referencing involves grading attainment on a standardised test against a known distribution of scores from the population of interest. Individuals are therefore evaluated directly against a reference group, as explained by Isaacs et al. (2013, p. 97):

> *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*

The adoption of norm-referencing would therefore be a significant paradigm shift towards a more explicitly psychometric approach to assessment. It demands a much more standardised approach to testing because the assessment content and difficulty

must be the same for each individual student in each examination series so that valid comparisons to the overall population (the 'norm') can be made. This would likely require significant use of common (anchor) items between examination series, which would make the security of the assessment of utmost importance.

Before considering each stage of the lifecycle, it is worth noting that many aspects of the current system would be unchanged under a norm-referencing approach to setting and maintaining standards, while others might change considerably. For example, it would still be necessary to hold examinations each year, which would then need to be marked and quality assured. Coursework, on the other hand, would not be possible because of the challenges around standardising the tasks (discussed later in this section). Difficulties would therefore arise with constructs that require a practical assessment format.

This report focuses on the potential differences to the current system; some fundamental, some more subtle. With that in mind, we begin by discussing some of the major technical and policy decisions that may be required to facilitate a norm-referencing approach. A large part of the standard setting happens in a study to establish the results for a normed population of interest.

5.1 Establishing the 'norm'

5.1.1 Who is the population of interest?

A fundamentally important task would be to establish the 'norm' that acts as the reference for each examination series. This exercise would be partly technical but would also require informed policy decisions. First, it would be necessary to define the population of interest, essentially specifying who is in the reference group against which the norm will be established. Some questions that would need to be addressed include:

a. What is the age range of the population? (e.g. all 16-year-olds in Wales?)
b. Is the assessment intended to span the full range of possible attainment?
c. Is the composition of the population something which is to be reviewed over time?
d. If so, what will be the process for this and how frequently will reviews take place?
e. Will different populations be specified for each individual qualification (reflecting differences in uptake in, for example, English Literature or history)?

Reflecting on point 'e', frequent research ('norming' studies) would be required for each individual qualification in the GCSE suite. Such ongoing research would require considerable resource and could take a variety of forms. Indeed, inter-subject

comparability would be an important consideration, with the approach to individual subjects having implications for stakeholder engagement and public confidence.

Under norm-referencing, the process for awarding grades would differ significantly from that used in the current system. Grade boundaries would usually remain static each year – a product of the established distribution of scores in the population, combined with prior decisions about where the level of attainment for each grade should be set in relation to this distribution. Indeed, a decision about the proportion of students in the population who should achieve each grade would be required, along with decisions about the desired statistical characteristics of the reference distribution. For example:

a. Should the density of grades in the reference group be normally distributed (with the middle grade being the most likely to be achieved by a student who is selected at random), or should the proportion of students achieving each grade be more uniform (with all grades being equally likely, if selecting a student at random)?
b. Put another way, should the range of marks into which each grade is allocated be of a uniform width (such as 10 marks per grade) or will grade widths vary at different points along the distribution?
c. If non-normal mark distributions in the population are to be statistically transformed, what does this do to the representation of standards (for example, if the population is bunched at the bottom or top of the distribution before transformation)?
d. And, following on from the above, to what extent would distributions be transformed such that they were 'standardised' (in a statistical sense) across different qualifications?

As Lenhard et al. (2019) point out regarding psychometric tests that utilise norm-referencing approaches:

> *Many psychometric tests are based on the assumption 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.*
>
> (p. 2)

This core principle of a norm-referenced approach, that the test is evaluating a student on a (latent) construct which we expect to be normally distributed throughout the population, would need to be kept at the forefront of technical thinking during the qualification development phase, and indeed throughout the lifecycle. Adjustments would need to be made where there was evidence (e.g. from ongoing norming studies) that assumptions about the population were not being adequately met.

Many of the decisions outlined above will depend on the purpose of the qualification and the extent to which grades need to discriminate between the attainment of candidates. GCSEs are considered to have multiple purposes but are often used to differentiate between individuals based on their likely ability. For example, employers often use qualification grades to help them when deciding which candidates to interview for a job. Educators often use grades to select the best applicants to enrol on a competitive course. If differentiation is prioritised, it may be that top grades (A* and A) should be achieved only by a very low proportion of the candidature, thus representing exceptional achievement and better assisting employers and educators in their decision making. Of course, there are all sorts of implications around such policy decisions, particularly given the multiple purposes to which qualification outcomes are put (Newton, 2007).

5.1.2 How will the standard be referenced over time?

Norm-referencing would not introduce predetermined quotas of students for each grade. If a given cohort were to perform particularly well or particularly badly in relation to that established norm, the profile of grades for that cohort would reflect this. This could, potentially, lead to significant changes in outcomes between years, something that may have mixed implications for public confidence. Accusations of grades being suppressed by statistics, which are sometimes levelled at the current system, would be unlikely, but the public may question the reliability or legitimacy of the system if national outcomes were to fluctuate substantially in ways that defied explanation or were perceived to be unfair (e.g. if teaching and learning were significantly disrupted).

Stakeholders would need to understand that, as is arguably the case with the current system, the grade that a student received would not provide precise information regarding their knowledge and skills (as would be the goal of criterion-referencing). A grade would represent an individual's attainment relative to the population rather than their attainment in relation to the curriculum. For example, a grade B might represent attainment within the 25th and 10th percentile of the population. Whether students had attained a particular standard in, for example, algebra in their maths GCSE, would not be explicitly referenced.

Nonetheless, there may be ways in which typical performance standards at each grade could be captured and communicated to stakeholders. As is the case now, grade descriptors could be established that broadly described the level of attainment which students who achieve that grade tend to demonstrate through their performance, though these descriptors would need to be at a high level to allow for the many different routes through which a student may achieve their overall mark. Item level data from the assessments themselves could also be used to build up a profile to describe typical attainment (or performance) at each grade. Such profiles may provide some support to teachers and stakeholders, but it is important to stress that descriptions would be necessarily vague given that the assessment standard would not, in a fundamental sense, reference the curriculum directly.

When discussing the curriculum, it is important to note that the standardised assessment at the heart of a norm-referenced approach to standards would be highly vulnerable to curriculum change. Norm-referencing places central importance on the consistency of the assessment, such that it is the same each year in terms of content and difficulty. This differs from the current approach, which uses a combination of statistical and judgmental evidence to account, as far as is possible, for changes in assessment difficulty. Qualification reform, or even relatively minor changes to the curriculum, would present a significant challenge in terms of comparability between cohorts and may therefore undermine the validity of the approach. This is because, essentially, new cohorts would be being referenced against a population that had studied different content (at least to some extent) and taken a different assessment (at least to some degree). The system would be vulnerable to challenges on the grounds of its fairness, unless a high degree of standardisation could be achieved.

Returning to an earlier point, much may depend on how confident stakeholders were that the 'norm' against which test-takers were being graded was sufficiently representative and contemporary. Significant consideration would need to be given to the methodology and frequency of 'norming' studies and the degree to which standardisation was achieved without compromising a sense of fairness. As noted in the definitions section, research into the most frequently used example of norm-referencing, intelligence testing, illustrates how outcomes can increase over time (the 'Flynn' effect; see Flynn, 1987), a phenomenon which likely results from a combination of socio-environmental and measurement effects. Norming studies would be necessary to understand, and probably adjust for, such effects to maintain the integrity of the assessment over time.

Two potential frameworks for norming studies are briefly described below:

1. Regularly scheduled (e.g. biennial) adjustments to the distribution and/or grade boundaries. Such adjustments would be based on analysis of item level data obtained from each new cohort. This would mean that the 'norm' would be updated regularly to include or account for new cohorts as they are assessed, with any changes based upon performance on 'anchor' items that were common across all examination series (and had therefore been taken by the entire population). Though the flexibility of being able to make regular refinements is appealing, any change to the distribution could undermine the validity of the qualification, perhaps causing the assessment to drift from the construct(s) that it is intended to assess, or perhaps through the erosion of public confidence (e.g. 'the test is meant to be same for everyone, but it was harder when I took it').

2. Alternatively, a national assessment that sits alongside, but is separate to, GCSE qualifications could be established with the sole purpose of monitoring and maintaining the reference distribution and its associated grade boundaries. Such an assessment would be similar to the National Reference Test (Burge & Benson, 2022) that is used in England to support the awarding of their current GCSEs. The main advantage would be that the 'norming' studies would be independent of the qualifications. However, such an approach would come with considerable financial costs given the need to develop and deliver regular additional national assessments. Decisions would also be required about which subject(s) to include and at what scale. In addition, the assessments would be low stakes for the students taking them, which may undermine its validity as a proxy for the population norm. Moreover, any changes over time in student test-taking motivation would undermine the study design.

5.2 Would norm-referencing require significant changes to the assessment model?

Norm-referencing is usually used for setting and maintaining the standard of a single test, so if a modular assessment model was desirable (with subjects divided into separately assessed components), policy makers and assessment developers would need to consider several complex issues. For example, is the standard intended to represent attainment at the end of the two-year course of study or, for some/all modules, does it represent a level of attainment at a particular stage? How would a student's performance across the modules be aggregated to determine their final qualification grade? It would be important to consider the technical effects of aggregation when considering how norm-referencing would work in a modular system, as well as the potential consequences, both intended and unintended, on teaching and learning (Baird et al., 2019). Such issues would likely need to be considered on a subject-by-subject basis to account for the unique content and context of each.

Regardless of the assessment model, for a truly norm-referenced approach to be realised there would need to be a higher degree of standardisation than is required in the current GCSE model. There are different approaches to achieving this, which means that the assessment model (the design of the tests and the items, the process by which components are aggregated, etc.) would be at least somewhat dependent on key structural decisions taken at the outset of any reform. Next, we briefly outline some of the broad methods through which individual performances could be 'linked' back to the reference distribution.

5.2.1 Common tests

If one were to adopt a purist approach to norm-referencing, then each student would take an identical test and be marked and graded consistently against established grade boundaries. Such an approach would be extremely difficult, perhaps impossible, to deliver in the context of GCSEs because of the challenges around keeping such a test secure and unpredictable. Were the assessment materials to be leaked to the public, all subsequent cohorts of students would have such an advantage that they could not be graded against the reference distribution and the entire test would need to be overhauled, invalidating the purpose of choosing a norm-referencing approach to begin with.

Even if security was broadly maintained, it seems extremely unlikely that students would not remember aspects of the test and share information with later cohorts, leading to rapid grade inflation. Riseing outcomes have been observed in norm-referenced testing in the US, especially in the context of school accountability (Linn, 2000; Linn et al., 1990; Shepard, 1990).

The approach would be particularly problematic should the use of coursework be desirable, as without changing the brief from year to year, it is likely that teachers would become increasingly adept at helping their students to achieve the highest possible marks. Indeed, there is evidence of this happening with controlled assessment and coursework in the past (Opposs, 2016). It is hard to envisage how coursework could feature in a norm-referenced system.

5.2.2 Common items

An alternative would be to vary the assessment in each series but maintain some common items that all students complete. Psychometric methods could then be used to equate the different iterations of the test using these 'anchor' items. Item Response Theory (IRT), which is used as the underpinning psychometric paradigm for many educational assessment systems around the world (Furr, 2021), could provide some solutions here. In broad terms, it takes an item-focused approach, modelling the

probability of a correct answer (or a given mark) as a function of the individual student's ability and the difficulty of the item (Furr, 2021). This approach provides a framework for constructing assessments, linking and equating tests, and evaluating how individual items have functioned.

Though this appears to be an elegant solution, the significant issues of security and predictability would remain because some form of anchoring would be required. This would likely take the form of common items, a set of questions that would need to cover a substantive proportion of the curriculum and be present in each examination series. It may also be necessary to develop a more focused approach to the development of items, because IRT requires the deployment of low tariff item types, such as selected response (e.g. multiple choice) items. A narrower range of available item types for these common items may have a negative impact on the range of skills that may be validly assessed within the qualification. Higher tariff items could be included in the live assessments, but given that these would probably not be appropriate for linking, and could not therefore be common items, there would be a threat to validity: the assessment of certain skills, perhaps those higher order skills associated with essay (performance) items, would not be being standardised. There would likely be similar issues around curriculum coverage for common versus non-common items.

While there would be very significant risks around security and predictability, generally, the above approaches could be employed to operationalise norm-referencing for GCSEs assessed by examination. However, it is difficult to see how they could be successfully employed in GCSEs with coursework assessment. There would therefore be implications for the design of GCSEs. Indeed, assessment models that would be consequences of such an approach being considered for the Assessment of Performance Unit monitoring tests were hotly contested (Goldstein, 1979, 1980) and ultimately rejected in the 1980s. Among other points, it was argued that the statistical models would drive test design rather than educational considerations of children's experience of the curriculum. Multiple choice tests, it was argued, were likely to be used, and this was considered a retrograde step for assessment in England.

5.3 Design phase

This stage of the qualification lifecycle focuses on establishing the core objectives of the qualification's assessment and designing assessment procedures that will validly measure attainment against them.

5.3.1 Subject and qualification criteria

As is the case under the current system, this stage of the qualification lifecycle would be important. In many ways, it would remain the same – each qualification would require a clear rationale, subject content would need to be specified and described in detail, appropriate assessment objectives would need to be established, and a scheme of assessment would need to be put in place. The last of these, the scheme of assessment, would depend on how data about the population (the norm to be referenced) was to be generated and maintained, something discussed above.

Though the point may seem obvious, it is easy to overlook the importance of ensuring that the assessment objectives are carefully aligned with the assessment model. Assessing, for example, a holistic and synoptic approach to analysing a given period of history may be more difficult to achieve if the assessment model requires extensive use of standardised multiple-choice items for anchoring purposes. There would therefore be significant subject level considerations, with some subjects requiring larger departures from their current schemes of assessment than others.

5.3.2 Sample Assessment Materials (SAMs) – question papers, mark schemes

We have discussed the differing assessment models that may be used under a norm-referencing approach. If a model that prioritises a large bank of selected response items for anchoring is used, some carefully curated examples could be provided within the SAMs, though these could not later be used for live assessment.

The point about predictability is also important to reiterate here. If the test materials and the items become overly familiar to staff or students, or, worse, the use of specific questions that are used as anchors becomes predictable, then outcomes will become inflated. Essentially, learners would be able to perform better than their peers in previous cohorts on the test due to more tailored preparation rather than superior attainment. This would need to be considered when developing SAMs and other exemplar materials.

5.4 Development phase

The annual rhythm for the development of the assessment would depend on the details of the approach that was adopted. For example, an IRT approach that relied on a large item bank would involve a somewhat different flow of item development, with the focus being on individual items rather than the structure of entire papers.

To summarise some of the discussion in the preceding sections:

- The tests themselves would need to remain static under a norm-referencing approach, with standards expected to be the same across each examination series. For an IRT style approach, much of the work to develop the item bank will have been completed up front. In the first year of the assessment this would be a very significant task, akin to writing multiple series worth of assessment. These would likely drive the norming study, with items being used to establish the distribution of attainment within the population of interest. Certainly, the initial outlay of time and resource would be very high, though subsequent years would likely require far less work. Maintenance of the item bank would be required, along with 'norming' studies to ensure an appropriate norm was being referenced for each cohort.
- Given that the emphasis would be on delivering a standardised test against established norms, it would be of utmost importance to ensure that the test had the same level of demand for all candidates.

5.5 Delivery phase

During the delivery phase, the assessment is operationalised for candidates. The developed assessments are undertaken in schools and marked and graded. Finally, each candidate receives their qualification outcome in the form of a mark and/or a grade.

5.5.1 Examinations

As discussed above, written examinations would need to be standardised given that the grade boundaries (established through research, the norming studies) would need to be fixed. Assessments would need to be carefully linked between series, as discussed above, which would likely change their structure and the type of items which were used.

5.5.2 Coursework

There would be immense challenges associated with establishing coursework within an assessment scheme whereby standards were set and maintained using norm-referencing. The level of standardisation needed would likely undermine the purpose of coursework. If successive cohorts became better at the assessment due to familiarity and coaching, as we have seen in many instances of qualifications that feature such assessments (Opposs, 2016), this would fuel steadily increasing outcomes which may be perceived as 'grade inflation' by many stakeholders. It is easy to imagine a scenario whereby the proportion of students receiving the highest grades increased each year because of rapidly and disproportionately (in relation to examination performance) rising coursework marks.

To resolve this, there may be viable assessment models which 'decouple' coursework components from the grade (e.g. they would be graded separately using a different model, with the students' subsequent grade also reported separately). This may or may not be desirable from a policy perspective, and it is noteworthy that the implications for assessment, teaching and learning would vary considerably between subjects.

5.5.3 Exam papers are delivered to centres

A key requirement for successful norm-referencing is that those taking the test have no (or at least very little) knowledge of its content. This means that test materials must be kept secure not just prior to each examination series but also on an ongoing basis. Sharing past papers would potentially give subsequent cohorts an advantage. Security of the test materials would entail not only keeping them confidential before the examinations were sat, but, ideally, also afterwards. Test papers would be collected at the end of the examination.

5.5.4 Marking

Much of the current examination processes would remain the same, although the stakes associated with the application of a consistent marking standard would be high. Shifts in the marking standard would undermine the maintenance of standards given that cut- scores (grade boundaries) would not normally move from one series to another. Heightened monitoring would likely be needed. Fortunately, the probable shift towards selected response items would increase marking consistency and reduce the required resource (perhaps even facilitating the use of automated marking). Markers would likely need to keep the content of the exams confidential, and the stakes associated with any 'leaks' could be high.

5.5.5 Grading

The awarding process would change significantly under norm-referencing. Examiner judgment would not be necessary. It is likely that the resource currently deployed to set grade boundaries would be used to conduct the statistical analyses outlined below prior to the delivery of results.

5.5.6 Post-results reviews and appeals

As with the current system, a norm-referencing approach would require a system through which teachers or students could request a review of marking should they believe there has been an error or an incorrect application of the mark scheme. Due to the requirements for security, it is unlikely that access to scripts, the question papers and mark schemes would be allowed. The system would be less transparent and there could be a consequent impact on trust.

5.6 Review phase

Typically, a great deal of statistical analysis is conducted in the review stage of a norm-referenced test to ensure that the items are functioning in line with the statistical assumptions of the psychometric model. This involves analyses of the fit of the items and the candidate scores to the overall model, any drift in item parameters over time and differential item functioning by subgroups of candidates taking the tests (e.g. by gender and ethnic group). Further analyses would likely be necessary to estimate the measurement error resulting from various aspects of the modelling. Where substantive issues arose, there may also be qualitative investigations relating to the design of the assessments and their marking schemes, or administration. An evaluation of the adequacy of the norming study would form part of a review under this model.

5.7 Summary of Chapter 5

- This chapter discusses how a norm-referencing approach to setting and maintaining standards could affect the lifecycle of GCSEs in Wales.
- Norm-referencing would require establishing a 'norm' or reference group against which students' performance would be evaluated.
- Grade boundaries would remain static each year based on the established distribution of scores in the population.
- Norm-referencing would introduce significant changes to the current system, including the need for standardised testing and potential challenges with curriculum changes.
- Stakeholders would need to understand that grades represent individuals' attainment relative to the population, not their attainment in relation to the curriculum.
- Regular norming studies would be required to maintain the integrity of the assessment over time and address potential changes in the population.
- Two potential frameworks for norming studies are discussed: regular adjustments to the distribution and grade boundaries based on item-level data or establishing a separate national assessment to monitor and maintain the reference distribution.
- Norm-referencing would require consideration of complex issues in a modular assessment model and a higher degree of standardisation than the current GCSE model.
- A purist approach with identical tests for each student would be challenging to implement due to security concerns.

6 The effects of criterion-referencing GCSE qualification standards on assessment processes

This section considers how a criterion-referencing approach might operate were it to be introduced for setting and maintaining standards for GCSEs in Wales. It considers the impact of the change at each stage of the qualification lifecycle, unpacking the potential risks and benefits that may be introduced. Criterion-referencing represents a set of processes. In different forms, it is used for high-stakes qualifications in France, Sweden and Queensland (Baird et al., 2018, p. 302). It is worth revisiting our working definition for criterion-referencing:

Definition Candidates receive grades that tell us whether they met predetermined performance criteria.

Method Suitably qualified subject matter experts define the performance criteria required for the award of a qualification and/or for particular grades. In GCSE, grade descriptors could be thought of as the performance criteria to be met. These also relate to attainment objectives. Standard setting is not, strictly, a distinct phase in a system in which criterion-referencing is used because the grades can be allocated directly by qualified assessors. Instead of standard setting, verification processes are applied to quality assure that all due procedures have been heeded in the assessment process. In some cases, moderation is used, in which case the assessment judgment is also quality assured.

In what follows, we set out the main features of three key variants of criterion-referenced assessment (Table 4). As with all assessments, variations in practice operate *within* these categories too. However, the main differences between them are in how the criteria are deemed to have been met, the assessment formats associated with them and the view of the assessor's task. Each approach is mainly derived from different practice contexts, which has also coloured the developments and ways of thinking.

Ultimately, we take the third, standards-referenced, approach as the most likely way that criterion-referencing could be designed into GCSEs in Wales because it most closely relates to current practices. The first two approaches – early criterion-referenced and competence-based – have structures that are unlikely to fit the culture of schooling and assessment for GCSEs in Wales. We outline them in Table 4 because there are important distinctions between them that can have radical effects on qualification design, operation and interpretation. Knowledge of the evolution of these approaches is key to considerations of the potential use of criterion-referencing for GCSEs in Wales because

each would have significant consequences for the experience of preparing and taking the tests – backwash effects upon teaching and learning.

Table 4 Three key variants of criterion-referencing

Variant	Performance criterion	Typical formats	Assessor task	Practice context
Early criterion-referenced	Meeting a mastery-based pass score	Multiple choice	Scoring	US, curriculum-based assessments
Competence-based	Meeting every criterion	Performance	Observing and evidence-checking (checklist)	Vocational assessments
Standards-referenced	Meeting a broad description	Academic	Holistic judgment	Higher education and school-based education

6.1 Early criterion-referenced tests

The term 'criterion-referenced testing' was first used by Glaser in the US in 1963 to define a form of testing that was already being discussed as an alternative to the then-dominant norm-referenced testing:

> *a student's score on a criterion-referenced measure provides explicit information as to what the individual can or cannot do. Criterion-referenced measures indicate the content of the behavioural repertory ... 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)

Glaser coined the term 'criterion-referenced' to frame a debate about how to design tests and arrive at a score that would provide test score users with information about what students could actually do, rather than their position with regard to others – a score with 'content-meaning' rather than 'normative-meaning' (Geisinger, 2021, p. 52). Geisinger provided a summary history of the thinking that preceded and followed

Glaser, with reference to how criterion-referenced testing might differ from the previously dominant norm-referenced testing.

6.2 Design, development and grading of criterion-referenced tests

In the US, the criterion-referenced testing movement evolved alongside developments in mastery learning and instructional technology. Indeed, the title of Glaser's (1963) influential essay is 'Instructional technology and the measurement of learning outcomes: Some questions'. Popham (2014), a strong advocate of criterion-referenced assessment, considered that the benefits emanated from its intrinsic link to programmes of learning:

> *An inherent assumption of criterion-referenced assessment, then, is that by articulating with sufficient clarity the nature of the curricular aims being assessed, and by building tests that enable us to measure whether individual students have achieved those aims to the desired level, we can teach students better. Criterion-referenced measurement, in every significant sense, is a measurement approach born of and preoccupied with instruction.*
>
> (para. 9)

This link to programmes of learning is an important one for our consideration of the possible use of criterion-referenced standards. Geisinger (2021) pointed out that if the purpose of the assessment is not to compare the student with other students but to provide assurance that the student has learned the necessary knowledge and skills, then it could be argued that

> *the best possible item would be one that no-one answers correctly before instruction and all answer correctly after instruction*
>
> (p. 52)

As such, criterion-referenced tests are often designed to confirm what students can do, rather than to differentiate between them.

Key differences between criterion-referenced and other sorts of test are the processes and concerns brought into play during test design and development. Steps must be taken to define the criteria to be assessed, ensure that the test covers all the required criteria, and that marking and grading systems ensure that a student cannot pass unless they have achieved all requirements. There is no sampling of the domain, the

assessment is a census of the domain content and there is no compensation – candidates must display knowledge and skills across the entire domain. The test specification needs to be very detailed, with prescriptions for individual items and how these should be combined and presented to the student (Popham, 1992, 1994). Passing scores may be set at test level (assuming a single criterion is being tested), or the test may be designed as a series of sub-tests, each designed to provide assurance that a particular criterion has been met. In both instances, a threshold passing score is often defined in advance (Hambleton et al., 1978). Sadler (1987) summarised how criterion-referenced testing had developed:

> *To date, the criterion-referenced testing movement (which is probably stronger in the United States than in any other country) has focused on (a) detailed objectives and specifications of domains of knowledge; (b) objective testing, using items of empirically determined power to assess specific achievements; and (c) a combination of measurement and numerical cut-offs for making grading decisions (including those for mastery or minimal competence).*
>
> (p. 192)

In terms of setting and maintaining standards, the focus is on judgmental methods that rely on the expertise of the judges. There are many methods available, some of them cognitively difficult for judges, and most of them resource-intensive. Geisinger's (2021) view was that the most viable approaches are those that involve panels of judges making judgments on test items or test difficulty overall; the two methods outlined by Geisinger are the well-known Angoff and Bookmark methods. In the first, the judges are supplied with all test items and for each item asked to estimate the percentage of minimally competent test takers who would get the item correct. In the second, the judges are given a 'book' of test items, arranged from least to most difficult using the statistics from students' performances. Judges are asked to mark the most difficult item that they think a minimally competent student would answer correctly. In both approaches, discussion and subsequent rounds of judgment may follow, and an agreed formula is used to 'average' the judgments and arrive at a passing score (see Cizek & Bunch (2007) for more detail).

6.2.1 Differentiation

The close link between criterion-referenced tests and instructional programmes is often portrayed as a strength that sets criterion-referencing apart from attainment- or norm-referencing. It allowed US states to closely define the intended curriculum. The purpose of the test, then, was to check that instruction had taken place, or, at best, to provide an indication of the quality of the instruction.

If the assessment also serves other purposes, such as selection, the lack of differentiation can be problematic. These problems persist in derivatives of criterion-referenced tests today: in criterion-referenced assessment systems, especially ones that allow assessment when ready and re-assessment, results tend to be bunched at the top end of the distribution. Such an assessment design is useful when the main purpose of the assessment is to confirm that the school is delivering learning programmes effectively, but it is less useful as an individual qualification in education systems where the assessment is used to manage scarce progression opportunities. It can also create communication problems in educational cultures where social expectations are that qualifications should differentiate. Even if used only for some components of assessment, this bunching of results can be problematic: as we have seen in our discussion of coursework components in the current GCSE system, the result can be that assessments do not achieve their intended weight and risks can be introduced to the setting of overall qualification standards.

6.2.2 Overly technical specifications and processes

Criterion-referenced tests were designed to produce assessments that were more meaningful to teachers and students by being more directly related to their learning programmes. In practice, the complexity of the assessment design spawned an enormous industry and technical literature. Ironically, in moving in this direction, the criterion-referenced testing movement worked against its own aims, so that by 1987, supporters of criterion-referencing like Sadler were highly critical of how criterion-referenced tests had developed, claiming that despite their initial links to instructional aims, they had made assessment less transparent and moved it further away from the control or understanding of the teacher or the student:

> *It relies on relatively sophisticated statistical and technological solutions to the problem of grading students according to their actual achievements. Because of the complexity of the procedures, much of the responsibility for grading and assessment is removed from the teaching profession as a whole and vested within a central bureau or agency …. the processes leading up to grading, and the interpretation of the grades themselves, become less accessible to the teacher, the student and the private citizen.*
>
> (Sadler, 1987, p. 192)

This ‘sophisticated statistical and technological’ literature was summarised by Geisinger, who set out the thinking on appropriate ways to measure test reliability, judge content validity, evaluate item quality and determine test length (Geisinger, 2021;

or for a more contemporaneous review of the technical literature, see Hambleton et al., 1978).

Geisinger summarised different ways to set standards in criterion-referenced tests, but a fuller summary was given by Hambleton et al. (2000). Crucially, they were writing at a time when the limitations of criterion-referenced testing to capture complex skills had been realised, and test developers had started to use test items that Hambleton et al. call 'performance' items, that is, items that attempt to assess skills through a constructed response and more open-ended tasks requiring polytomous scoring rubrics. Almost all the standard-setting methods outlined are administratively complex for the testing organisation, cognitively complex for the judges, and time-consuming. Whether they could be applied in a national school-leaving system where assessments must be marked and graded and results issued within the space of a few weeks is questionable. It is likely that bespoke technological solutions would be required. As judgmental methods, they are all subject to the caveats about awarder judgment that we have outlined in our discussion of approaches in current GCSEs. Next, we turn to a different way of viewing criterion-referencing: competence-based assessment. This approach is not associated with multiple-choice tests or sophisticated statistical models underpinning standard setting. Rather, it is much more closely tied to written criteria.

6.3 Competence-based assessment

In the UK, criterion-referenced assessment has largely been developed in the field of vocational and occupational qualifications. The term that tended to be used was 'competence-based' assessment. SQA qualifications use a competence-based approach for some of their unit-based qualifications (for example see SQA, 2017). Wolf (1995), the foremost critic of that assessment approach, defined competence-based assessment as a 'specialised development' (p. 3) of criterion-referenced assessment. We will follow Wolf's lead in maintaining this distinction.

6.3.1 Design, development and grading of competence-based assessments

Competence-based education and assessment is widely accepted as having originated in the US, in the field of occupational training, particularly teacher training (Tuxworth, 1989; Wolf, 1995). The term 'competence' is a key one and denotes a strong link to the needs of industry and the economy. The ethos of competence-based education was also rooted in a wish to democratise training, both in its availability and in how it was delivered. Jessup (1989a), the chief architect of competence-based assessment programmes in the UK, emphasised this aspect:

By specifying the competences sought independently of the learning process, access to learning through any mode becomes possible. Along with unit credits and credit

accumulation, continuing education and training will be made available to sectors of the population which have never participated in the formal system.
(p. ix)

Tuxworth (1989), setting out the history of competence-based programmes in the US, summarises a set of features that appear to place the learner at the centre:

1. Individualisation of learning
2. Feedback to learners
3. Emphasis on exit rather than admission requirements
4. Systematic programme
5. Modularisation
6. Student and programme accountability

(p. 15)

In the UK, several government White Papers throughout the 1980s increasingly advocated this approach as an answer to the lack of coordination and consistency across the country's vocational education and training system. In 1986 the National Council for Vocational Qualifications (NCVQ) was established to oversee a comprehensive programme of 'standards development' aimed at overhauling the system (Jessup, 1989a; Wolf, 1995). For those in charge of implementing the new approach, their aim was to transform what they saw as out-of-date programmes of learning, inappropriate assessment methods, and an inflexible vocational training system that excluded many people from entering and made progression difficult for those who had been able to embark on learning Jessup (1989b). The new qualifications were called National Vocational Qualifications (NVQs):

NVQs require an explicit statement of competence, that is, a specification written down for everybody to see, in an agreed and recognisable format. The statement of competence spells out what candidates are required to be able to do for the award of an NVQ, and includes the criteria by which performance can be assessed. In doing so, the statement of competence also sets clear goals for educational and training programmes.
(p. 68)

Like competence-based education in the US, NVQs were intended to expand learning opportunities by widening modes of learning, that is, the places where learning could happen: not just in formal education settings but also in workplaces and through open learning. Perhaps because of this, the intention was that assessment requirements should be seen as open and flexible, not prescriptive:

The statement of competence is also independent of the method of assessment. The nature of the competence will indicate the category of evidence required for assessment (i.e. performance demonstrations, knowledge, etc,) but within that category options will exist as to the specific method or instrument of assessment adopted. This legitimises other forms of assessment such as assessment in the workplace and assessment of prior achievement, as well as assessment by more conventional methods. These are all seen as alternative forms of evidence of competence.
(Jessup, 1989b, p. 71)

Despite these early laudable aims, competence-based qualifications in the UK, particularly in England, have been heavily criticised. The bodies responsible, the names and features of the system, and the individual qualifications have all been reviewed and reformed many times. Despite this, NVQs continue to be offered and to be the main occupational qualification offered in all the UK nations. They also retain their basic design of being competence-based, and outcome-focused, with assessment requirements that derive from explicit performance criteria.

An example of a current competence-based qualification is the Pearson (2017) Level 2 NVQ Diploma in Housekeeping. It is designed for people working in a wide range of hospitality and catering businesses. It comprises 3 mandatory units and a choice from 12 optional units. One of these optional units is entitled *Collect Linen and Make Beds*. It is a Level 1 unit, requiring 21 guided learning hours and includes the following four learning outcomes:

1. *Be able to collect clean linen and bed coverings.*
2. *Understand how to collect clean linen and bed coverings.*
3. *Be able to strip and make beds.*
4. *Know how to strip and make beds.*

Each learning outcome addresses knowledge and skill separately – 'understand how…' and 'be able to…' respectively. Each learning outcome has between 4 and 7 assessment criteria. For example, the assessment criteria for the first learning outcome are as follows:

1. *Choose and collect the linen and bed coverings needed for work schedule.*
2. *Make sure the linen and bed coverings meet organisational standards.*
3. *Handle and move the linen and bed coverings safely.*
4. *Keep linen store safe and secure.*

To be awarded the NVQ certificate, a learner will need to demonstrate competence across every assessment criterion, for each learning outcome, within all units taken.

6.3.2 Over-specification of requirements

For the early designers of the UK's competence-based qualifications, transparency in the standards to be achieved was a key aim. However, how *can* textual descriptions be explicitly and unambiguously meaningful to the range of assessors who will use them?

Wolf described how, over just a few years of design and development activity, NVQ specifications were required to become increasingly detailed. When statements of standards that consisted of outcomes and performance criteria failed to provide clarity, qualification designers were required to add range statements, then to add descriptions of underpinning knowledge and understanding, then evidence indicators, and then, since none of that appeared to give clarity and specificity, 'Amplification' was required.

This was done in a vain attempt to obtain complete clarity and a set of criteria that could be unambiguously used to assess competence in an occupational role (Wolf, 1995, pp. 21–27). Wolf blamed this failure on an insistence on a one-to-one relationship between criteria and competence, and between criteria and assessment (see Wolf (1995, chapter 4, and 1996) for multiple examples). Rather than clarity, elaborate over-specification resulted.

6.3.3 Inconsistent interpretation

For the early designers of NVQs like Jessop, direct observation and judgment of a range of evidence by a professional was intended to be a mode of assessment that was more open, flexible and fit for purpose. Jessop saw such 'evidence-checking' as sitting alongside more conventional assessment methods. For critics of the system, though, the new system was not more flexible, but less. For Wolf, for example, a corollary of the detailed nature of competence-based assessment specifications was that there could be no practical way to check the candidate's evidence of competence except by direct observation. It was assumed that detail and clarity in the criteria made the task less one of judgment and more one of the professional or occupational expert simply recognising the appropriate performance (Wolf, 1995, p. 27).

However, expert judgment of how evidence matched a specification proved to be less than straightforward. Citing several evaluations of competence-based assessment in the UK, Wolf concluded that assessors, even when relatively expert in their field, found consistent interpretation of detailed assessment criteria to be a difficult task (Wolf, 2001, pp. 9–10). For example, a report by Black et al. (1989) documented and analysed four case studies of National Certificate assessment in further education colleges in Scotland. The research found low levels of agreement on assessment decisions between

lecturers in the same college, lecturers in different colleges using the same assessment instrument, and between lecturers and researchers. Their conclusion was that problems arose from several factors, including:

> *the need to rely on the professional judgment of lecturers, the problems which arise from inadequate domain sampling, [and] the flexibility in the decision making procedures which encourages spurious borderline decisions.*
>
> (p. 81)

6.3.4 Context-free assessment

Performance criteria were also designed to be context-free (especially in transferable skills like communication or broad occupational areas like administration), but this also affected the consistency of assessor judgment. Wolf noted:

The inherent variability of the contexts in which competence is tested and displayed means that assessors have to make constant and major decisions about how to take account of that context when judging whether an observed piece of evidence 'fits' a defined criterion.
(Wolf, 2001, p. 9)

Assessors will always have to judge skills within a specific context and, given that no set of criteria can be tailored for every possible context (indeed, they are often designed to be generic), assessor judgment cannot involve objective comparison but instead represents a subjective interpretation of the criteria (Allais, 2012; Cresswell & Houston, 1991). This problem affects judgments of similar skills that cross occupational boundaries – are communication requirements the same for hairdressers as they are for bricklayers, for example? It also affects judgments of skills within an occupation. In the Black et al. study, in some cases assessors reported that they consciously interpreted the criteria differently for different students (Black et al., 1989). From a teaching point of view, they quite rightly took account of the student's starting point, their prior knowledge and experience; from an assessment point of view, of course this renders the assessment inconsistent and unreliable.

6.4 Standards-referenced assessment

Sadler distinguished between 'criterion-referenced' assessment and 'standards-referenced' assessment (Sadler, 1987). Standards-referenced assessment is his attempt to report students' actual achievements while increasing the possibilities for school-

based assessment and acknowledging the centrality of human judgment in the assessment process.

For Sadler, objective testing is inadequate for assessing skills in subjects that require a judgment of quality (Sadler, 1987, p. 191). He noted subjects where there may not be a correct or incorrect response – 'English and other languages, visual and performing arts, manual and practical arts, humanities, and social sciences' (Sadler, 1987, p. 193). For Sadler, meaningful assessment of student achievement should not be a process of binary classification but one in which the teacher judges the student's performance against a series of criteria, envisaged as existing on an underlying continuum of quality.

In promoting the place of teacher judgment, Sadler (1987) was concerned to make explicit the 'tacit standards' by which they may have made such judgments in the past. For him, standards, seen as 'a fixed point of reference for assessing individual students' (p. 191) must be made visible to both teacher and students (and potentially the wider public). Grading is done by the teacher, who makes a holistic overall judgment of learner proficiency, based on a pattern of performance over a series of tasks or assessments.

Accepting evidence of inadequacies in teacher judgments, Sadler proposed that those judgments are made dependable through well-expressed criteria and additional tools and training. The criteria should be text-based descriptions of acceptable performance, written in natural language, accompanied by explanatory materials in the form of exemplars of student responses. Teachers may be given opportunities to come to mutual understanding of the standard through discussion of the criteria and exemplars (Sadler, 1987).

Sadler's definition of standards-referenced assessment is arguably the variant of criterion-referencing that has had most influence and use in general and academic assessment and qualifications. Indeed, he mentions practice in Queensland, which along with practice in Scotland and New Zealand is perhaps globally the most well-established use of criterion-referencing approaches in the assessment of general, curriculum-based school qualifications. Writing about practice in Queensland for the New Zealand Council for Educational Research, Klenowski and Wyatt-Smith (2010) expanded on Sadler's ideas about how to support teacher judgment, arguing that it is only by discussing examples of student performance that teachers can come to an understanding of the standards to be assessed. Hence, for Klenowski and Wyatt-Smith, social moderation systems are crucial in criterion-based teacher assessment systems (p. 114).

Klenowski and Wyatt-Smith (2010) pointed to several advantages of social moderation: it requires teachers to make the basis of their judgments explicit, improving inter-rater

reliability and providing a quality assurance check that can increase confidence in the judgments (pp. 114-118). By increasing teacher understanding of the standards to be achieved, social moderation is also felt to have a direct benefit for the quality of teaching and learning in the classroom (pp. 117-118). They also stress the importance of the community of practice, seeing this as a route to social construction of knowledge about the standards and the curriculum. In this model, teachers are not passive recipients of curricula and standards but active creators of them (pp. 119-120).

6.5 What might a criterion-referenced GCSE qualification look like?

Before considering each stage of the lifecycle, it is worth noting that many aspects of the current system could be unchanged under a criterion-referenced approach, whichever variant or variants of that approach was used. In practice, many decisions would be required to define how the system would work in context, and each of these decisions would have implications for the design, development and delivery of the qualifications. The following section therefore takes a rather broad view of how the system might be operationalised and is not intended to provide an exhaustive breakdown of how it would work in practice.

In a criterion-referenced system, the assessment structure could be very similar to the current one, that is, a combination of classwork or coursework, perhaps set and judged by teachers, and examinations, set and judged by an external body, with criteria, perhaps in the form of grade criteria, being used to steer assessment and grading decisions, either at the level of individual tasks, or at the overall qualification level, or both. Changes to the assessment formats are not necessarily a feature of a move to criterion-referencing.

We follow Sadler in assuming that a pure or strict form of criterion-referencing, close to either criterion-referenced tests or competence-based assessment, would be very difficult, if not impossible, to implement in large-scale, high-stakes end-of-school qualifications. In a report for Ofqual, Burdett et al. (2013, p. 13) noted the following:

> *Strictly speaking, a true criterion-referenced system would test whether every individual is (or is not) capable of achieving each and every one of the criteria contained within the curriculum framework.*

In their study of standards maintenance system in 10 jurisdictions, the system in New Zealand was found to be closest to strict criterion-referencing, but the authors noted that

> *[i]n practice, in large-scale end of schooling assessment, it is not possible to test every criterion within the curriculum framework.*
>
> (p. 13)

For our depiction of how criterion-referenced standards might operate in GCSEs, we will adopt a model that is based on Sadler's conception of standards-referenced assessment, with some features drawn from other concepts to make things concrete. As such, the model that we will depict has the following features:

- Performance is judged against a series of grade criteria, envisaged as existing on an underlying continuum of quality.
- These criteria take the form of text-based descriptions of acceptable performance, defined in advance by subject experts.
- The criteria provide more detail than traditional grade criteria, but there is not a one-to-one relationship between criteria and assessment tasks.
- These criteria are used in the process of coming to all assessment judgments, albeit they may be used at different points and in different ways for different components of the assessment.
- The criteria can be used as part of the learning process, to help students to understand their own progress.
- The criteria do not define the standard on their own. They are supplemented by contextualised and annotated student responses which are expected to play a key role in standard setting.
- Grading comprises holistic judgment of learner proficiency, based on a pattern of performances taken over a series of tasks or assessments.
- Because grading relies on an extensive pattern of performance, teacher assessment of classroom and coursework tasks plays a role in determining the final assessment result.
- Teachers/assessors are given support through appropriate tools and training and are part of a community of practice.
- Social moderation systems form the basis for communities of practice that allow teachers/assessors to come to a mutual understanding of the standard through discussion of a combination of verbal descriptions and exemplars of student responses.

Of course, this broad description of a criterion-referenced approach leaves many aspects of policy and technical design undefined. Like any approach to standard setting, a criterion-referenced approach has inherent drawbacks: it would not be possible to design a system to totally overcome these flaws. We will discuss the threats to standards that would be generated in such a system, but it is important to note that the

features outlined above, singly or in combination, form part of the more long-lasting, accepted, or even successful criterion-referenced grading systems. Documenting the form of all such systems is beyond the scope of this paper. We now turn to the likely effects of such an approach to standard setting on the qualification lifecycle.

6.6 Design and development phase

In a criterion-referenced system this is arguably one of the most significant stages in the qualifications lifecycle, and one that can be at the centre of debate and controversy.

6.6.1 Qualification design, specifications and assessment criteria

At a high level, the requirements of this stage would remain the same as in the current system. Each qualification would require a clear rationale, subject content would need to be specified and described in detail, and appropriate assessment objectives would need to be established.

For example, in the SQA (2017) National 4 Applications of Mathematics course, the Course Specification outlines the rationale for the course, both generally, in terms of its relationship to the principles of the curriculum, and specifically, in terms of the skills, knowledge and aptitudes in mathematics that will be developed. This course is modular, consisting of four units, and the specification provides a summary of the aims of each of the units, with an emphasis on developing skills and knowledge in real-life contexts through practical applications of skills that develop individual confidence. In doing so, the specification sets the standard within the context of a set of wider curricular and societal purposes. The aims of the unit in Numeracy (National 4) are described as follows:

> *The general aim of this Unit is to develop learners' numerical and information handling skills to solve straightforward, real-life problems involving number, money, time and measurement. As learners tackle real-life problems, they will decide what numeracy skills to use and how to apply these skills to an appropriate level of accuracy. Learners will also interpret graphical data and use their knowledge and understanding of probability to identify solutions to straightforward real-life problems involving money, time and measurement. Learners will use their solutions to make and explain decisions.*
>
> (p. 6)

As in the current system, there would need to be a scheme of assessment that sets out the high-level assessment design principles for this specific qualification at that level.

As outlined above, criterion-referenced approaches are linked to concepts of mastery (Glaser, 1963) and arose from a movement that sought to check that learning had been achieved rather than to distinguish who had and had not learned (Popham & Husek, 1969). As such, the scheme of assessment is likely to specify that at least some of the assessments should be taken when the student is ready. Assessment is likely to be continuous rather than taking place at endpoints (Wolf, 1996, p. 220). Re-taking an assessment will probably be allowed, and rather than prescriptions on re-sitting, the scheme of assessment is likely to contain guidance on how much and what sort of assistance from the teacher is acceptable between attempts at assessment tasks.

As in the current system, design principles may specify the number and content of assessment tasks, the form of the assessments for each unit, any constraints around the length of assessments and any conditions for assessment, such as restrictions on the use of aids. Before this, though, subject criteria would need to be developed and agreed:

Since criterion-referenced tests are specifically designed to provide information that is directly interpretable in terms of specified performance standards, this means that performance standards must be established prior to test construction

(Glaser & Nitko, 1971, p. 654)

Front-loading of the standard setting process engenders the nub of the issues with criterion-referenced standards. Transparency of standards is fostered for assessors, teachers, students and the public (Popham, 1993). The approach can be attractive to policymakers for that reason (Baird, 2007, p. 138). However, practical and theoretical questions are raised. In an early exploration of the possibilities of using a type of criterion-referenced system in UK school qualifications, the notion of domain-referencing was explored. This is similar to criterion-referencing but a different term was used to delineate a broader use of criteria in a compensatory manner, with sampling of the domain. Reporting on investigations into the possible use of such an approach, Christie and Forrest (1981) suggested that

> *[a]lmost all … syllabuses and their associated examinations have undergone a long and careful process of consensual validation: they do not refer to what any one person chose to teach and measure.*
>
> *To this extent public examinations in England and Wales are already domain-referenced.*

(pp. 28-29)

Such easy consensus among experts, if it ever existed, is probably something that many current assessment professionals long for. In the 21st century, what constitutes knowledge, what constitutes a subject and what constitutes a domain or subdomain of a subject tends to occasion fierce debate within and across subject communities. Policymakers themselves may also have deeply held views about particular subjects. In such a context, we need clear policy steers on who gets to sit around the table when the decisions are made - and what is the process by which the criteria to be used are arrived at? Decisions would need to be taken on:

a. Who specifies the criteria to be assessed?
b. On what basis do they do so? What model are they asked to use, and what processes are they asked to follow?
c. How is disagreement to be resolved?
d. What design features will be common across subjects and domains?

These questions have been at the heart of criticisms of a number of embodiments of a criterion-referenced approach, and the policy decisions are not trivial but may be questions on which the system stands or falls.

6.6.1.1 Qualification design, specifications and assessment criteria – threats to standards

As much of standard setting is front-loaded, what is specified in the subject content and specifications is of prime importance to subject experts keen to see their conception of their subject (or sometimes of cross- or interdisciplinary aspects of learning) included in the assessment specification. Such upfront specification of assessment requirements may be perceived as controlling what is assessed and therefore what is taught, and, as Christie and Forrest (1981) noted, even in the more consensual times that they inhabited, 'there is likely to be a much more active interest taken in the credentials of subject committees' (p. 56). The expertise of those who define the criteria is central to this approach to standard setting, with its emphasis on authentic assessment.

An example of this was exhibited during the Curriculum for Excellence development programme in Scotland. Between 2004 and 2014, an outcomes-based curriculum was developed, intended to transform all learning from ages 3 to 18 and beyond; of course, this necessitated major reform of the qualifications system. In some subjects, like history, subject specialists hold differing views on both the subject content and the fundamental conceptions of the purpose of the subject and therefore the key skills to be taught. This became obvious in public and private statements made by an influential group of history academics (The Newsroom, 2011).

Arguably, this example illustrates what happens when one group of experts feel excluded from the process of writing subject content and defining outcomes and criteria. For the experts who have a seat at the table, the issue remains of how consensus is to be reached. When Linn (1995) was discussing progress towards defining national content standards as part of Bill Clinton's *Goals 2000* reforms, he pointed to the difficulties of achieving consensus on what should be taught and assessed. He noted:

> *Consensus clearly becomes more difficult to achieve as curriculum materials and assessments make standards more concrete and specific.*
>
> (p. 13)

Since the point of assessment or grading criteria is precisely to make standards concrete and explicit, we should not be surprised to learn that consensus is difficult to achieve. Citing examples of controversy over history standards, Linn (1995, p. 14) concluded that

some level of controversy over content standards is inevitable. The struggle over what gets emphasised, what gets included, and what gets excluded from the content standards, performance standards, and assessments is a struggle over educational values.

Linn also noted that while differences of opinion on the key content and concepts to be taught and assessed are to be expected among subject specialists, differences of opinion can also occur between subject specialists and the public, or between subject specialists and advocates of cross-disciplinary and interdisciplinary initiatives. Differences of opinion can occur between the school educationalists involved in writing the criteria and the university academics who will receive holders of the qualifications. For example, in the early 1980s a panel overseeing admission to mathematics courses in Scottish universities complained about the draft criteria for the new Standard Grade in Mathematics. It was designed to be a criterion-referenced qualification for 16-year-olds. Subject content had been reduced too much, basic techniques and skills had been omitted, and knowledge had been neglected in favour of problem-solving skills (Philip, 2009, p. 186): all important themes that we find in debates about subject qualifications today and of yesteryear. Such disagreements are not unique to a criterion-referenced system, but they are more material when knowledge and skills must be specified upfront.

Processes of qualification design often happen behind closed doors, and it is difficult to find documentary evidence of the qualification designers' reactions to such expert differences of opinion. Professional experience suggests that it is reasonable to expect that in many instances, assessment professionals and policymakers alike, rather than resolving conflict, or even making a decision on one side or another, find themselves simply trying to include everyone's 'vital' content. This contributes to one of the key risks to standards of criterion-referenced assessment, that of elaborate over-specification.

Policymakers must accept that precision, and therefore strict criterion-referencing, is not desirable or possible. Popham stressed that criteria should be pitched at a 'mid-level of detail' (Popham, 1992) and proposed that we seek to develop 'Goldilocks' domain descriptions, in which the level of descriptive detail is neither too brief nor too elaborate, but just right' (Popham, 2014). In contested fields, this becomes highly political.

There is not a straightforward solution. Rather, the literature tells us that the answer may lie in a combination of factors, including accepting that assessor judgment has a larger part to play, albeit in the full knowledge of its limitations. Sadler summarised many of the criticisms of teacher judgment and the reasons cited for its fallibility, and for him, these failings happened because studies were looking at how well teachers grade intuitively (Sadler, 1987, p. 194). Since then, though, there have been a plethora of robust studies on teacher judgment that have illuminated circumstances in which that judgment can be compromised (Baird, 2007; Urhahne & Wijnia, 2021). Nevertheless, since judgment must play a part in a criterion-referenced system, policymakers would need to find ways to mitigate the risks associated with it.

6.7 Delivery phase

During the delivery phase, the assessment is operationalised for candidates. The assessments are undertaken by students and marked and graded. Finally, each candidate receives their qualification outcome in the form of a mark and/or a grade.

6.7.1 Setting the assessment and mark scheme/judging criteria

Teacher assessment often plays a much larger part in criterion-referenced systems and might take a number of forms that go beyond what we would normally class as coursework, such as observations of performances or portfolios. It is feasible for examinations to be part of the assessment schemes, and for examination development to mirror the sorts of processes and to have the same issues and concerns as in the current system. Here, we will highlight only areas where practice may differ under a criterion-referenced system.

6.7.1.1 Setting the assessment and mark scheme/judging criteria – teacher assessment

Other than in the criterion-referenced testing movement in the US, there has long been a close association between criterion-referenced assessment and teacher assessment. There are several reasons for this. Sadler (1987), for example, argues that so-called educational 'measurement' in fact consists of many micro-judgments made by teachers, with scores assigned as proxies for those judgments. Measurement, he argued, is simply a way of coding and combining the sorts of judgments that teachers make of their students every day (p. 193). Why not, then, use a more direct system and try to use and improve the teacher judgments?

In the assessment systems surveyed by Burdett et al. (2013), several (e.g. in Alberta, Hong Kong and New Zealand) were found to make use of teacher assessment and were often linked to a concern with explicit standards. Two of the key aims of teacher assessment that often overlap with the aims of criterion-referencing are that performance can be judged over an extended period and that assessment can be linked more directly and explicitly to the aims of the curriculum.

There are many different models of how teacher assessment could operate. Teachers have a lot of flexibility in some systems. At the other extreme might be a system of teacher assessment that would fit within the high levels of specification previously defined by UK regulators for 'controlled assessment' (see Opposs, 2016). The assessment could be set by an external body or conducted in very strictly controlled conditions (like an examination), or the outputs of the assessment could be marked or judged by examiners who are external to the student's school or college. If all three of these conditions apply, the assessment would not be teacher assessment but might be more accurately termed 'coursework', a term currently in use.

Perhaps more typical of a criterion-referenced system might be an assessment system that looks more like the reformed system in Queensland, Australia. Here, after more than 40 years of a criterion-referenced, wholly teacher-assessed system, demands for reform have resulted in a hybrid system (see Cumming (2020) for a history of the system and analysis of the drivers for reform). In summary, in the reformed qualifications, assessment for senior school courses comprises four summative assessments per course, usually with 75% of the final subject result coming from teacher-assessed components. Based on pre-defined standards, the QCAA (Queensland Curriculum and Assessment Authority) see their system as providing 'a balanced, integrated assessment programme'. They describe the defining characteristics as shown in Box 4.

Teacher training, guidance and a quality assurance system that supports while ensuring quality in teacher-developed assessment tasks is crucial for criterion-referenced assessment. This support can take different forms. For example, for many years SQA has produced and disseminated a *Guide to Assessment*, (SQA, 2019b) to help teachers and assessors to devise, administer, mark and grade assessment tasks and assessment instruments that will meet requirements for validity and reliability.

To return to the Queensland example, assessment literacy and integrity are seen as key aspects of teacher professionalism. Involvement in QCAA processes is seen as important staff development for teachers, and the claim is made that,

> *The system invests in teacher knowledge and expertise and fosters a culture that trusts and empowers them to do their work.*

Heavy involvement in curriculum development and assessment processes are promoted as 'develop[ing] teachers' pedagogical practice and assessment literacy' (QCAA, 2023b).

Box 4 Queensland Certificate of Education

Evidence of student achievement is gathered over time from a range of complementary approaches to assessment that have been selected because of their relevance to the purpose of the assessment and to the knowledge, skills and understanding to be assessed. Assessment techniques include projects, investigations, extended responses, performances, products and examinations.

The validity of assessment is improved by assembling evidence of student achievement from a variety of assessment techniques and conditions. Reliability of assessment is improved by providing students with multiple opportunities to demonstrate their knowledge, understanding and skills, as well as by collecting evidence at different times and under different conditions. Accessibility of assessment is achieved through measures such as ensuring all students have a clear understanding of how to demonstrate their learning, considering accessibility of language and layout when developing assessments, and implementing appropriate principal-reported or QCAA [Queensland Curriculum and Assessment Authority]-approved access arrangements and reasonable adjustments.

The QCE [Queensland Certificate of Education] system is based on an innovative model of senior assessment that combines the flexibility and authenticity of school-based assessment, developed, and marked by classroom teachers, with the rigour and consistency of external assessment set and marked by QCAA-trained assessment writers and markers.

For decades, Queensland teachers have been reporting student achievement based on evidence collected from school-based assessment. This is an important consequence of valuing different techniques of assessment and seeking to provide teachers with meaningful professional development that improves their assessment skills and expertise. School-based assessment requirements are described in the syllabus, with guidelines for teachers on the conditions and techniques for assessment. Particular assessment approaches are mandated, but the syllabuses also allow teachers to contextualise assessments to the particular characteristics of the school and students. School-based assessment is marked by classroom teachers using advice in syllabuses.

(QCAA, 2023b)

6.7.2 Setting the assessment and mark scheme/judging criteria – threats to standards

6.7.2.1 Issues of interpretation

For critics and supporters of criterion-referenced assessment alike, the problem of interpretation is perhaps the most difficult one to deal with, and this holds true in all variants of criterion-referenced assessment. We have already noted the imprecision of language as a communication tool, and there are numerous studies that document the difficulties that 'experts' have agreeing key terms and definitions when attempting to define competences (for example, Markowitsch and Luomi-Messerer (2008) provide a striking account of the difficulties of arriving at European Qualifications Framework criteria). Discussing the post-apartheid outcomes-based curriculum in South Africa, Allais (2012) described an example from language courses:

For example, an outcome such as 'show an awareness of manipulative devices' can be displayed by primary school children (e.g. through nursery rhymes), by newly literate adults (e.g. through understanding of simple slogans) and by people using language for a high level of academic proficiency.
(p. 342)

Orr and Forrest (1984) in an exploratory study designed to inform the development of criterion-referenced GCSEs, found that experienced examiners struggled to make consistent judgments on the skills covered by test items and how these related to assessment objectives. This issue persists in current uses of criterion-referenced assessment. Take the Standard and Testing Agency's national curriculum assessment requirements: in English language, one requirement is 'exercise an assured and conscious control over levels of formality, particularly through manipulating grammar and vocabulary to achieve this' (STA, 2018, p. 5). Without seeing this in the context of the other requirements, including the levels above and below, it would be impossible to tell that this is a requirement for those judged to be working at greater depth than the expected standard in Key Stage 2 (ages 7–11).

6.7.3 Teacher assessment is conducted

In the current system, as we have seen, there are rules in place around the way in which non-exam assessment is conducted, and similar rules are equally likely to be needed and used in a criterion-referenced system and to vary according to the subject and task. Checks on the authenticity of the work will be needed, as well as some level of control to ensure that demands placed upon students are consistent. Some tasks may be completed by students with very few limitations on the task set or the environment in which the task is taken.

In a criterion-referenced system, it is likely that we would continue to see this range and variation in ways that teacher assessment can be conducted. For example, in Queensland, the requirements for a 'Collection of work' are defined like this:

> *A collection of work assesses a response to a series of tasks relating to a single topic in a module of work. The student response consists of a collection of at least three assessable components provided at different times and may be demonstrated in different circumstances and places.*
>
> (QCAA, 2023c)

By contrast, when carrying out a practical demonstration, the student is required to respond individually and in a set timeframe. This range of ways of conducting teacher assessment is typical in criterion-referenced systems, which generates issues of comparability.

6.7.4 Examinations are conducted

In most instances, if examinations are part of a criterion-referenced system, how they are conducted will not differ from how this is done in the current system. An emphasis on authenticity of assessment may result in differences in permitted examination conditions, for example what resources the student is allowed to take into the examination room, but these sorts of differences are not peculiar to criterion-referenced systems.

If responsibility for setting and conducting examinations is devolved to the school or college, as in SQA Higher National Qualifications and Queensland Applied subjects, then very similar processes and rules are likely to apply: the responsibility for ensuring that the rules are followed sits with a different group/organisation, but the rules are likely to be very similar to the rules for conducting examinations in any other system. For example, SQA (2019a) guidance on graded units sets out the requirements:

> *Supervision may be carried out by a member of the course team or by external individuals contracted by the centre. The management of the examination is the responsibility of the centre and it is recommended that all aspects should be carried out by a clearly identified person from each centre, e.g. the examination co-ordinator or SQA co-ordinator. The roles and responsibilities of supervisors will include:*
>
> - *receipt and security of examination papers at the examination event*
> - *distribution of examination papers to learners*

- *overseeing examinations to ensure that examination regulations and conditions are complied with*
- *reporting back, especially where incidents of malpractice occur*
- *collecting learners' papers and returning them to the examination co-ordinator/SQA co-ordinator or other named person.*

(p. 24)

6.7.4.1 Teacher assessment is conducted – threats to standards

6.7.4.1.1 Teacher and student workload

The issues of interpretation that we have already outlined, the potential flexibility in the form and timing of assessment, and the possibility of assessment when ready and re-assessment are all designed to bring positive benefits. Together, though, they bring a major threat to standards, and one that has more than once hastened the breakdown of criterion-referenced assessment systems, which are the implications for teacher and student workload.

In theory, an assessment system that is designed to be flexible and able to be tailored to the learning programme, and indeed to the individual learner, is one that should be least stressful for student and teacher. In practice, this often proves not to be the case. Gathering evidence of skills and knowledge for whole class groups can be a time-consuming activity for the teacher. It is sometimes argued that, at best, the result is a checklist or 'tick-list' form of assessment (Sadler, 1987, p. 195; Wolf, 1996, p. 221). This is especially the case in systems that use detailed criteria, where mastery of each has to be evidenced, but, as with other risks to standards of criterion-referenced assessment, it remains a risk.

Teachers are often exhorted to make holistic judgments based on a range of evidence. For example, the Queensland handbook quoted above noted that 'Reliability of assessment is improved by […] collecting evidence at different times and under different conditions' (QCAA, 2023b, para. 9). Take the Queensland General Senior Syllabus for English (QCAA, 2019). This subject is split into four units. For each of the first two units, the teacher is required to develop and administer formative assessment tasks. For Units 3 and 4, the teacher must administer three summative assessments and mark or judge these. In this subject, the three teacher-assessed components of summative assessment are all extended responses (two written, one spoken). The teacher is provided with Instrument Specific Marking Guides which describe the qualities of work required to meet the assessment objectives. While in practice there may be little difference between a criterion-referenced Instrument Specific Marking Guide and the sort of levels-of-response mark scheme that is commonly used in the

current GCSE marking system, in Queensland the class teacher will have to mark at least three of these for every student in the class.

Workload is not just a problem for the teacher. For the student, it is not just the assessment workload in one subject that is relevant, but the total workload across all of the subjects that they are taking. Notwithstanding, students typically favour assessment systems that are not 'all or nothing' in a final exam, even when they recognise the faults in such systems and whether or not they are criterion-referenced (see, for example, Barrance & Elwood, 2018, p. 259 or SQA, 2016, p. 39).

6.7.4.1.2 Consistent approaches to learner support

As outlined in our discussion of the current system, when teacher assessment is conducted there is a risk that schools and colleges come to different interpretations of the amount of assistance teachers can give to students, and this inconsistency can be difficult to detect through any moderation or quality assurance processes. Arguably, this threat is exacerbated in a criterion-referenced system because teacher support and re-assessment are often seen as acceptable processes in such systems. For the awarding organisation, the threat to standards might be dealt with by producing and promoting guidance on when and how support can and should be given. For example, in its guidance on Higher National Graded Units, the SQA (2019a) defines and exemplifies the concept of 'reasonable assistance' while noting that the concept requires a degree of assessor judgment as to what might be appropriate in any given situation (pp. 17-18).

6.7.4.1.3 Quality assurance of teacher assessment

Mature criterion-referenced assessment systems recognise that they cannot sidestep problems of interpretation and instead must find a way to deal with it. This is often done by introducing various sorts of quality assurance regarding how the assessment is set, conducted and judged in the school or centre. In Queensland, the quality of assessment is assured through a two-stage process. In the first stage, 'Endorsement', the school must submit its assessment instruments for internal assessment to be checked by an external, independent subject expert, who evaluates the validity and accessibility of the draft assessment instrument against the following criteria:

- opportunities for students to demonstrate relevant subject matter and assessment objectives
- opportunities to demonstrate the range of performance levels/syllabus standards
- alignment to assessment specifications for the technique
- conventions for item construction
- scope and scale of the assessment items for the defined syllabus conditions
- authentication strategies for the assessment instrument

- scaffolding that informs students about the requirements for their response
- language and layout for the technique and intended audience.

(QCAA, 2023a)

In the second stage of quality assurance, 'Confirmation', schools are required to submit a defined sample of student work to QCAA. QCAA assessors review this selected sample of student responses to summative internal assessments for every subject in every school to check the accuracy and consistency of teacher marking of their students' work. The sample selected varies from subject to subject and school to school, depending on a number of factors, including the school's previous Confirmation results. Similarly, the corrective actions taken by QCAA may vary. These corrective actions may involve recalibration of all of the cohort's marks by QCAA officers 'supported by a rules-based algorithm', or a requirement that the school must re-mark all of the work of that subject cohort (QCAA, 2023d).

We have seen that for QCAA one important facet of ensuring assessment quality lies in an emphasis on teacher professionalism, supported by a framework of teacher guidance and training. For proponents of criterion-referenced assessment, systems of support for teacher assessment professionalism are key to the success of the assessment system and provide the answers to ensuring consistent interpretation of criteria. While we have noted the potential inadequacies of teacher judgment, critics of criterion-referenced assessment have themselves provided some pointers to ways in which judgment may be strengthened. For Cresswell, judgments about student achievements and standards are always value judgments (see, for example, Cresswell, 1996). Central to achieving quality in these value judgments is the experience and knowledge of the judges, but Cresswell also sets out a vision of an alternative model of the judgment process, in which judges engage 'in a constant process of evaluation and re-evaluation as they read the candidate's work' (Cresswell, 2000, p. 81).

Foreshadowing development of social moderation practices, Black et al. (1989) observed that 'the most notable example of high comparability occurred in the study in which we encountered the greatest amount of collaboration amongst staff':

The Communication lecturers were aware that they were dealing with an ill-defined area and to have any chance of comparability between lecturers (and between colleges) they would have to consult one another. By doing this they could reach agreement about the meaning of the learning outcomes and performance criteria in their subject area. This takes a lot of time and effort, however, and not all the problems have been solved; but this area, which has the least precisely defined descriptor of any in our case-studies, and contains the greatest scope for subjective judgment, has produced the greatest consistency in decision-making.

(p. 80)

In this respect, achieving consistency of assessment is based not on the clarity of the criteria, or even on finding a 'Goldilocks' mean between specificity and manageability, but on providing appropriate support in the form of exemplars of student performance, guidance materials, training opportunities, and, crucially, opportunities for assessors to discuss and agree on standards. Teachers need to be involved in discussions about what is being assessed, why it is being assessed and how it will be assessed. Social moderation in which groups of teachers take responsibility for standardising and quality assuring assessment judgments can become communities of practice that support development of practice and expertise (Hutchinson & Hayward, 2005).

6.7.5 *Malpractice and maladministration*

In many ways, issues around malpractice and maladministration in a criterion-referenced system are similar to those in the current system. Teaching to the test has often been seen as more problematical due to the overt specification in criterion-referencing. This is only deemed malpractice under certain circumstances in which students are overly supported and the rationale for the qualification is undermined. However, lack of teaching in criterion-referenced systems has also been criticised. Torrance (2007) deemed this 'assessment as learning', in which students experience little or no instruction but are simply assessed against the competences. After all, the work to be done on the course has been made explicit. Both of these issues affect the validity and therefore standards of a qualification.

6.7.5.1 *Grading and issuing of results*

Whether criterion-referenced or not, qualification systems have to define how the results of assessment components will be combined. There are well-established ways to do this in systems that use numerical marking. Where the judgments may involve direct grading, or comparison with criteria, this can be more problematic. Decisions have to be taken about how much information is to be conveyed by the final result and how meaningful this information needs to be. There are two main ways to combine the results of individual assessment components, and they have different implications for the uses and meaning of those results.

6.7.5.2 *Aggregation*

The first way to combine results is to aggregate them in some way, as traditional marking systems do. If the design of the assessment task has used marks to value student evidence, those marks can be combined numerically and converted into a grade. This could mirror the current GCSE system. If the design of the assessment task has required the assessor to make direct grading judgments, using letter grades or descriptive terms like 'Merit' or 'Distinction', the most typical way to do this would be

to assign a numerical value to those non-numerical data and combine the numerical values using agreed rules of combination. If there are not too many distinct grades, an alternative might be to create a matrix or look-up table showing each possible combination of grading values and the resultant overall grade to be awarded.

6.7.5.3 Profiling

The second way to deal with the need to report achievement of a number of different criteria is to report the results in a profile. In its most direct form, the profile would simply list the assessment criteria that the student had mastered. This is simple (if time-consuming) to produce. In practice, most criterion-referenced profiling systems have involved rules of aggregation for individual assessment criteria, with only a small number of sub-domains reported for each domain or subject. This may be done through a modular or unitised system, as in the current system in New Zealand, but need not be. For example, in Standard Grade qualifications in Scotland, first introduced in 1983, each subject was reported in the form of a profile of 'elements'. There were usually two or three of these for each subject. These were also combined into an overall grade for the subject.

6.7.5.3.1 Aggregation – threats to standards

The issue of how to aggregate attainment across different assessment components can be a problem for designers of criterion-referenced systems. With a system that relies on detailed criteria, it is difficult to specify the weight to be attached to each attribute defined as necessary (Cresswell (1996, p. 65) summarises findings originally published by Wilmut and Rose (1989)). After all, not all aspects of the curriculum are equally important. Another issue, and one that is important for perceptions of fairness, is how the rules of combination make allowances for uneven performance across different aspects of the domain. Such issues are most problematic in systems that attempt to combine the achievement of detailed criteria (Cresswell, 1987, 1988, 1994).

6.7.5.3.2 Profiling – threats to standards

The main issue with profiling achievement is how useful and understandable the profile is to the users of the qualification. Detailed profiles can be uninformative for qualification users. In the UK there have been several attempts to use profile reporting of assessment results (e.g. Records of Achievement; see Fairbairn (1988) or Hart et al. (2010, pp. 24–25) for a summary of their implementation), and most have been discontinued because they have been viewed as either too long to be helpful or too couched in 'assessment speak' to be understandable. Also, because in most public examination systems a primary purpose of the qualification is to allow selection for further study and employment, an overall grade is almost always retained alongside the profile report. Cresswell (1987, 1988) details the reasons why this renders either the profile or the overall grade or both of them less meaningful and/or reliable.

6.7.6 Post-results reviews and appeals

The likely reliance on teacher assessment rather than on examination would necessitate a system in which students can challenge their teachers' judgments. This is possible in the current system, but the demand is relatively low, and the nature of coursework is that there is usually an artefact produced which can be re-evaluated. Reviews may involve procedural checks that all of the necessary processes have been followed. Less likely in these systems is a re-evaluation of the assessor's judgments, as moderation or verification has usually already taken place and assumed to quality assure the standards.

6.8 Review phase

As evidenced in this chapter, a great deal of research has been conducted internationally on criterion-referenced assessments. In a formal review process, as indicated previously, there would likely be a discussion of the criteria and assessment principles, and a review phase would entail finding ways to deal with the differences of professional opinion that we documented while discussing the initial design phase. Equally, there would likely be a great deal of tension regarding assessment reliability, given what we know about the limitations of judgments and criteria. Reliance on social or consensus moderation in order to standardise teachers' interpretation and application of assessment standards may make the efficacy of these processes a subject of micro-validation, as it has been in Australia (Klenowski & Wyatt-Smith, 2014).

Perhaps the most important issue for the review phase is when it is allowed to occur and what is within the scope of the review. Major reforms of qualifications and assessment happen rarely (Isaacs & Gorgen, 2018) and are difficult and costly to implement, even if widely supported. For most systems, moving school qualifications to a criterion-referenced system will involve a paradigm shift in professional and public thinking about assessment, and there will be at least a few years of disagreement about whether the shift is the right one to make. In the early years, practical difficulties can seem insurmountable, and this can lead to a recurring cycle of demands for major reform. The system needs a degree of consensus and strong-willed policymakers willing to see through initial difficulties, or it can get caught in a vicious circle of review that simply waters down the aims of the initial reform, and in doing so, risks losing the best aspects of that reform.

In this section we have illustrated the benefits and problems that could occur in the qualification lifecycle with the adoption of criterion-referencing. Importantly, GCSEs were originally conceived as criterion-referenced qualifications, but the approach had to be adapted and essentially failed in its pure form (see Box 5).

A re-marking exercise using examination scripts further compounded the problems as the criteria proved ambiguous and bore little relation to the actual responses of candidates. At this point the draft grade criteria were dropped – except for the WJEC English GCSE.

The working group for GCSE English had a different approach, producing far fewer, and broader, criteria. This made differentiation between performances at different grades difficult. For example, to get a grade A/B in the writing domain, candidates should 'Give a coherent and perceptive account of both actual and imagined experience', whereas for grade F/G they should 'Give a coherent account of personal experience'. However, this broad-based approach was better aligned with established examination assessment practice. This approach was further developed by WJEC, which issued a 1988 GCSE English syllabus which incorporated some of the criterion-referenced approach.

The organisation of the 1988 WJEC GCSE syllabus for English would be recognised today by its stated *aims* and *objectives*, with candidates required to demonstrate their ability to meet the eight objectives (WJEC, 1988). It made clear that 'it is neither desirable or practicable to identify elements of the assessment within which specific objectives will be tested in isolation'. The evidence that this was only loosely criterion-referenced came in the introduction to the grade descriptions:

> *The grade award will depend in practice upon the extent to which the candidate has met the assessment objectives overall and it might conceal weakness in one aspect of the examination which is balanced by above average performance in some other.*
>
> (p. 8)

This acknowledgement of compensation within the awarding process distances this kind of approach from strict criterion-referencing but made grading more feasible.

Box 5 Criterion-referenced approaches in GCSE

In 1984, the then Minister of Education, Sir Keith Joseph, announced,

> *Examination grades should have a clearer meaning and pupils and teachers need clearer goals. We accordingly need grade-related criteria which will specify the knowledge, understanding and skills expected for the award of particular grades.*
>
> (DES, 1987)

In preparing for the introduction of the GCSE in 1988, there was an aspiration to produce a criterion-referenced qualification which would generate direct information about the capabilities of candidates. This was in line with the attempts to make National Curriculum assessment criterion-referenced, with the initial policy intention to incorporate the GCSE within it.

There were already *grade descriptions* in the GCSE subject criteria that gave a broad idea of the level of performance likely to have been shown for a particular grade. The Schools Examination Council (SEC) set up working parties in each of the main subjects to produce more informative descriptions of candidate performances. These working parties first identified within-subject *domains* which were then broken down into abilities, incorporating criteria to be met at each level.

The draft grade criteria were put out for consultation in 1985. Feedback showed that the working parties had produced an unmanageable volume of criteria. For example, history had three domains with ten sub-elements across four levels of performance leading to forty statements of performance. The mathematics working group formulated eighty detailed criteria – for one domain, at one level. This fuelled the fear that criterion-referencing would affect teaching and learning as a result of

> *very tightly defined syllabuses and patterns of assessment which would not allow the flexibility of approach that characterises education in this country.*
>
> (SEC, 1984, p. 2)

6.9 Summary of Chapter 6

- This chapter examines how the introduction of a criterion-referencing approach could affect setting standards for GCSE in Wales.
- Criterion-referencing involves grading based on predetermined text-based performance criteria.
- Three key variants of criterion-referenced assessment are discussed: early criterion-referenced, competence-based and standards-referenced. The standards-referenced approach is considered the best fit for GCSEs in Wales.
- Standards-referenced assessment relies on a broad description and requires holistic judgment. Criterion-referenced tests are designed to confirm what students can do rather than differentiate between them.
- The assessment structure in a criterion-referenced system could be similar to the current system, including classwork, coursework and examinations, with grade criteria used for assessment and grading decisions.
- Teachers and assessors are supported through training and belong to a community of practice, while social moderation systems foster understanding of standards through discussions.
- Specifying assessment criteria involves establishing performance standards, determining who sets the criteria and what model they are based on, resolving disagreements and deciding common design features.
- Elaborate over-specification of criteria is a risk in criterion-referenced assessment, and policymakers must find a balance between precision and inclusiveness.
- Teacher judgment plays a part in criterion-referenced systems, and policymakers need to mitigate the associated risks. Assessment reliability and the use of social or consensus moderation to standardise teachers' interpretation and application of assessment standards are important considerations.
- Interpretation issues arise in criterion-referenced assessment due to imprecision of language and difficulties in defining competences.
- Teacher and student workloads pose threats to standards in criterion-referenced assessment systems.
- The GCSE qualification in the UK initially aimed to be criterion-referenced but had to be adapted due to practical challenges.

7 Looking ahead to new GCSEs based on *Curriculum for Wales*

7.1 Background

The introduction of the new Curriculum for Wales marks a significant milestone in the evolution of the Welsh education system and will impact assessment in Wales. *Curriculum for Wales* is the cornerstone of the Welsh Government's efforts to reform education in Wales and build an education system that raises educational standards and enjoys public confidence. Under the reforms, each school is developing its own curriculum, supported by national guidance. As a result, several reforms are taking place. These are aimed at ensuring that the assessment process better reflects the needs and priorities of Wales and provides students with qualifications that are more closely aligned with their experiences and the Welsh education system. Implementation of the reforms began in 2022 in schools for age groups up to Year 6 and Year 7. New GCSEs will be introduced for first teaching in September 2025 and will be awarded for the first time in 2027. Some of the key reforms to GCSEs include:

1. Wales-specific qualifications: GCSEs are being developed specifically for Wales, with syllabuses and assessment criteria that reflect the future needs and priorities of Wales to complement existing assessments available across the UK and internationally.
2. Decoupling from the English system: The Welsh Government has continued to decouple the GCSE and A-level systems in Wales from those in England so that the assessment process is more closely aligned with the Welsh education system. This includes increased flexibility regarding both the content that can be taught and the focus of the assessment.
3. Wales-specific assessment criteria: The assessment criteria for GCSEs and A-levels in Wales have been revised to better reflect the needs and priorities of Wales and to provide students with qualifications that are more closely aligned with their experiences and the Welsh education system. Requirements for assessment formats such as examinations and coursework will differ between Wales and England.
4. Modular structures and greater use of teacher assessment: The continued use of modular exam structures, which allow students to take their exams in smaller, more manageable parts, as opposed to linear exams that test a full range of knowledge and understanding in one sitting. This feature aims to support student learning and provide opportunities for students to demonstrate their achievement. The nature of the subject content lends itself to greater use of teacher assessment and many of the new GCSEs will have a greater proportion of non-exam assessment than is currently the case.

5. Use of digital technology: Using technology to broaden the range of evidence that can contribute to a student's grade, to help make teacher-led assessment more manageable and to make the assessment experience more relevant and engaging.
6. Focus on well-being and mental health: The introduction of new measures to support student well-being and mental health during the examination process, including the use of a mix of different assessment methods within a qualification to cater for different preferences, and the provision of additional support and resources for students who may experience stress or anxiety.

(Qualifications Wales, 2021a)

Wales is reforming its approach to school accountability – seeking to move away from accountability structures built around high-stakes performative measures. Rather, schools will be expected to use qualifications data to self-evaluate and improve (Welsh Government, 2022).

7.2 Features of the reformed GCSEs and their implications for standard setting

The changes to the curriculum and the associated reforms to GCSE will have implications for how standards are set and maintained (Qualifications Wales, 2023b). The intention is that the new GCSEs will offer relevant, authentic and engaging assessment, with sufficient flexibility to allow schools to design their individual curricula. This, and the nature of the subject content, require more use of teacher assessment rather than exams in many subjects. Many of the new GCSEs will have a greater proportion (often more than 50%) of non-exam assessment than is currently the case. Further, a greater range of subjects will include onscreen assessment, and it is expected that this will be more engaging for learners, allowing them to respond to a greater range of stimulus material. Digital technology is also planned to support practical assessments, making them more manageable and authentic. Other recent technological advances, such as those in generative artificial intelligence (e.g. Bard and ChatGPT), will also need to be considered given their potential to have indirect (and unintended) consequences for assessment.

Decisions regarding the standard setting approach are yet to be taken, though it has been decided that the standards and outcomes will be broadly similar to those for the current set of GCSEs (Qualifications Wales, 2023b, pp. 35-36). A comparable outcomes approach was adopted for the first awards of recent reformed qualifications (though note that this was more like attainment-referencing in Wales). Candidates receive, as a group, comparable qualification grade outcomes to those which they would have received had they followed the course before a reform and taken the old qualification

(see page 11 for a fuller description). This is to protect candidates from a drop in outcomes experienced because they happen to be among the first students to take the new qualifications with which teachers are less familiar.

Implementing such an approach in modular qualifications is challenging (Baird et al., 2019). Early unit awards are made without knowledge of how they will aggregate into overall qualification outcomes. This is affected by the way in which the results combine across the assessments, which is in turn caused, at least in part, by the way students have been prepared and entered for the units. There is a degree of uncertainty about these early unit awards, as the ability and preparation of early cohorts is somewhat unknown. Hence, there is an emphasis on examiner judgment of performance in setting grade boundaries despite the content and assessments being quite different to that of legacy qualifications. Further, examiner judgment cannot account for unknown aggregation effects, which impact overall qualification outcomes. Setting grade boundaries for early non-exam assessment units can also be tricky as teachers may expect that boundaries will not change in future series, and once grade boundaries are set, teachers use them to inform their teaching. These difficulties, however, have been mostly well managed in past reforms.

Comparable outcomes may give way to a more strongly attainment-referenced approach after the first few years of awards. Either way, the maintenance of qualification standards can be challenging in qualifications with a large proportion of teacher assessment. Marks for teacher-assessed units often increase year on year, as teachers become more familiar with the demands of the tasks and more exemplars and support materials are available. As discussed above, penalising students for *lack of familiarity* with a new qualification is not normally seen as warranted, so rewarding students *for* familiarity would be perverse (Cuff et al., 2019). Notwithstanding, awarding committees are sometimes reticent about increasing grade boundaries in line with these increases in marks because doing so disrupts relations between teachers and students. In contrast, marks for examined units increase only a little as familiarity with the qualification increases. To maintain standards at the overall qualification level, the grade boundaries on the examined units are typically increased to compensate for changes in marks on the teacher-assessed units. This will affect the achieved weighting of the units in the overall qualification outcomes, undermining validity. Units with very high grade boundaries will contribute less to the overall grade than intended (for an explanation of achieved versus intended weights, see Adams & Murphy (1982)). Outcomes in teacher-assessed units can end up having little impact on overall qualification outcomes if they differentiate insufficiently between students.

If the mark distributions for examined and teacher-assessed units significantly diverge, it can be impossible to satisfactorily maintain standards at qualification level. There may

be year-on-year increases in outcomes driven by improvements in performance on teacher assessments but not matched by improved performance in examined units. These increases may undermine confidence in the grades. To maintain the overall outcomes, boundary marks may be increased on the written examinations to compensate for the increases in coursework outcomes. This can mean that the boundary marks become too compressed. With few marks between grades, reliability of grading is then threatened, which can also undermine confidence in the grading. In modular qualifications the standard setting is further complicated by the need to ensure comparability of standards across the routes through the qualification, with students taking modules at different times. The aim is that the standard required should be the same whatever the timing and sequence of the assessments taken.

In some qualifications these difficulties have led to a focus on maintaining standards at unit level without reference to the standards of the overall qualification. This is uncomfortable when users of qualification grades focus on the overall outcome not the unit grades. Further, it does not solve the problem of year-on-year increases in outcomes. The advantages are that it significantly simplifies standard setting, grade boundaries for examined units don't become compressed, and perhaps most importantly, it might be easier to ensure comparable standards across the routes through the qualification.

7.3 Are there features of criterion-referencing that are compatible with the new GCSEs?

The intention is that the new GCSEs will be sufficiently flexible to allow schools to design their individual curricula. It is hard to see how this is compatible with strict criterion-referencing, which would require very tightly defined specifications and assessments. *Curriculum for Wales* also seeks to promote an integrated approach to learning, which could be at odds with the specification and assessment of detailed criteria in isolation from each other. Further, the workload for teachers involved in assessing many, many criteria for each of their students would likely be prohibitive and distract from teaching.

While the meaning of grades is plainest under norm- or strong criterion-referenced standard setting approaches, it is still possible to generate a broad understanding of what is required to achieve grades in attainment-referencing. There may be elements of weaker forms of criterion-referencing that may be helpful in this regard and that could be incorporated into an attainment-referenced approach. For example, grade or performance descriptors, materials to support an understanding of performance standards, and opportunities to build a stronger community of practice around assessment standards.

Greatorex (2005) defined grade descriptions as 'indicators which exemplify the qualities candidates are likely to exhibit if they achieve a particular grade' (p. 9). They are a qualitative articulation of the skills and attributes associated with performance at the level of a particular grade. They provide an important link between examiners and teachers because they serve to make examiners' implicit understanding of performance standards explicit (Greatorex, 2002; Sadler, 1987).

There is debate about whether grade descriptors ought to reflect mid-grade performance or the minimum performance required for a grade. To be most useful in grade boundary setting amplification of minimum performance is preferable. But mid-grade performance descriptors are likely to be most useful for teachers and more reliable (Cadwallader, 2014). There is also the risk that in the early years of awarding, students' performance fails to reflect the grade descriptors. However, despite these complexities and risks, grade descriptors may be welcomed by teachers seeking to understand what is expected of students. It would be important to manage teachers' expectations regarding their likely precision, however. The possibility of forms of performance descriptor at unit as well as qualification level could also be explored.

In keeping with Sadler's views (1987), a number of other steps could support an understanding of the performance standards required. For example, SQA (SQA, n.d.) provides stakeholders with a wide range of subject-specific 'Understanding Standards' materials. These elucidate the standards required in the assessments, with examples of candidate assessment evidence. Such materials would complement grade descriptors, not least by demonstrating some of the many different routes to a grade in a compensatory qualification such as a GCSE. More comprehensive reports from Chairs of Examiners, setting out features of students' performance could form part of such a repository. Further, item-level data could be used to generate a profile to describe typical attainment (or performance) at each grade. This approach comes with some challenges, since exemplification is not available in the first year, when it is most crucial for teachers and learners. Examples need to be constructed and can be somewhat artificial. However, although constructed examples can illustrate the intended levels of performance, this may be unfairly high in the first years of the new qualifications, as teachers and learners are still becoming accustomed to the requirements.

Criterion-referenced systems rely on a strong community of assessment practice among teachers, with systems of social moderation playing a key role. While the current system of moderation by inspection in Welsh GCSEs is long established, there are elements of some models of social moderation which could helpfully support understanding of the content meaning of grades among teachers. More social opportunities to develop a consensus on standards and to clarify the performances that satisfactorily meet those standards could be established. These opportunities could cover both examination and

non-examination standards. In the latter case, more occasions for teachers to meet in groups (beyond their own school or college) to discuss their marking, to standardise their interpretation and application of the assessment standards, could be designed to not only improve the consistency of marking but to also improve the content meaning of grades. Moreover, there is evidence that this may also strengthen teachers' *Assessment for Learning* capability (Smaill, 2020).

Wholesale adoption of criterion-referencing for the *Curriculum for Wales* GCSE presents challenges for standards, since the approach is known to be less consistent than the more controlled and centralised procedures associated with examinations. Inconsistencies across teachers, schools and colleges, between years and between GCSEs in Wales and elsewhere are likely to arise. As discussed earlier in this report, moderation procedures could ameliorate these problems, but they are likely to remain to some extent. No approach can perfectly address standard setting: there are policy choices to be made regarding priorities.

7.4 Summary of Chapter 7

- The new *Curriculum for Wales* is being implemented to reform the education system and raise educational standards.
- New GCSEs will be introduced with Wales-specific qualifications, assessment criteria, and more emphasis on teacher-assessment and modular structures.
- The standard setting approach for the new GCSEs is yet to be determined, but comparable outcomes have been used previously to protect students from drops in grades due to unfamiliarity with new qualifications.
- Maintaining standards in qualifications with a large proportion of teacher assessment can be challenging, and adjustments may need to be made to grade boundaries to compensate for changes in marks awarded ion teacher-assessed units.
- Criterion-referencing may not be fully compatible with the flexible design of individual curricula in the new GCSEs, but elements such as grade descriptors could be incorporated to support understanding of performance standards.
- Wholesale adoption of criterion-referencing in the *Curriculum for Wales* GCSEs presents challenges for maintaining consistent standards across teachers, schools and years.
- Moderation procedures can help address inconsistencies, but some level of variation is likely to remain should criterion-referencing be considered a viable standard setting method.
- The effect of criterion-referencing on teachers' workload would need careful consideration.

7.5 Conclusion

Distinct approaches to standard setting have developed in response to evolving policy priorities in different contexts. We see this in the array of approaches that are taken internationally (Baird et al., 2018), as well as in changes in approach to the GCSE over time. As qualifications are necessarily culturally embedded, this is not surprising. Indeed, large changes in how standards are set are uncommon (Isaacs & Gorgen, 2018).

The impact of standard setting methods on disadvantaged students has not been researched per se. Of course, disadvantaged groups tend on average to have lower outcomes, but the selection of a particular standard setting method may not affect the rank order of students' outcomes; instead that depends upon the interaction between students' performances and the assessment criteria. *Where* the standards are set could reduce attainment gaps if a higher proportion of students are given higher grades, simply by reducing the discrimination between students' performances. If there is no need for fine discrimination, this could be seen as positive. However, the exact effect depends on where in the mark distribution the grade boundaries are set and the kinds of students whose work is found around those boundaries. Since GCSE outcomes are used in a variety of ways, it is likely that some purposes will continue to require fine discrimination. Also, it is noteworthy that even when policies were introduced in England in 2015 to make the GCSEs more rigorous, the effect of this on the attainment gap was ameliorated by the application of comparable outcomes (Burgess and Thomson, 2019). Due to the maintenance of an overall outcome profile for the population of GCSE-takers, attainment gaps were not much altered.

Assessment structures and formats may affect motivation and students' performances, although some research on GCSE outcomes found that modular assessment (Pinot de Moira et al., 2020) and teacher assessment (Pinot de Moira, 2020b) did not reduce attainment gaps. A wider literature review, however, did find evidence of bias in teacher assessment against disadvantaged groups (Wei Lee and Newton, 2021).

There is no perfect approach to standard setting. The way in which standards are embedded into qualifications plainly varies according to the approach adopted and each approach brings different threats to standards and to the valid interpretations of grades. It is notable, however, that we were unable to identify a norm-referenced qualification. Further, high stakes general qualifications do not lend themselves to strict criterion-referencing. This is perfectly illustrated by initial attempts to produce strongly criterion-referenced GCSEs. There are, however, elements of criterion-referencing which can support the content meaning of grades and give clarity to the curriculum aims being assessed to allow teachers to teach better. Recent decisions with respect to standard setting for the new suite of Made-for-Wales GCSEs indicate that different

approaches to standards may be taken once the new qualifications are established (Qualifications Wales, 2023b, p. 36). While it is anticipated that outcomes will be stable with the introduction of the new GCSEs, the approach may be adapted in the future to reflect changes in the performance of the population of learners in later years.

Given the broad nature of the current approach to standard setting, attainment-referencing, it may be possible to further integrate some helpful elements of criterion-referencing to strengthen the content meaning of grades. Post-pandemic, *Curriculum for Wales* represents an opportunity to revisit the approach taken to setting standards for GCSEs in Wales and to build a broad understanding of standards within the teaching profession. This report and the associated research programme provide supporting information for stakeholders involved in the standard setting for GCSEs in Wales policy considerations and communications.

8 References

Adams, R., M., & Murphy, R. (1982). The achieved weights of examination components. *Educational Studies*, *8*(1), 15–22.

Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 259–278. https://doi.org/10.1080/0969594X.2010.546775

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

AlphaPlus Consultancy Ltd. (2012). *The evaluation of the impact of changes to A levels and GCSEs*. Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/197437/DFE-RB203_1_.pdf

Baird, J.-A. (2007). Alternative conceptions of comparability. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms, *Techniques for monitoring the comparability of examination standards*. Qualifications and Curriculum Authority. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/487054/2007-comparability-exam-standards-f-chapter4.pdf

Baird, J.-A., Caro, D., Elliott, V., Masri, Y. E., Ingram, J., de Moira, A. P., Randhawa, A., Stobart, G., Meadows, M., Morin, C., & Taylor, R. (2019). *Examination reform: Impact of linear and modular examinations at GCSE* (OUCEA/19/1 Ofqual/19/6506/2). Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/798047/Modular_Linear_GCSE_summary_final.pdf

Baird, J.-A., Caro, D. H., & Hopfenbeck, T. N. (2016). Student perceptions of predictability of examination requirements and relationship with outcomes in high-stakes tests in Ireland. *Irish Educational Studies*, *35*(4), 361–379. https://doi.org/10.1080/03323315.2016.1227719

Baird, J.-A., Chamberlain, S., Meadows, M., Royal-Dawson, L., & Taylor, R. (2009). *Students' views of stretch and challenge in A-level examinations*. Assessment and Qualifications Alliance. https://filestore.aqa.org.uk/content/research/CERP-RP-JB-01032009.pdf

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., & Dhillon, D. (2005). *Qualitative expert judgements on examination standards: Valid, but inexact* (No. RPA_05_JB_RP_077). Assessment and Qualifications Alliance.

Baird, J.-A., Isaacs, T., Opposs, D., & Gray, L. (Eds.). (2018). *Examination Standards: How Measures and Meanings Differ Around the World*. UCL IOE Press.

Baird, J.-A., & Scharaschkin, A. (2002). Is the whole worth more than the sum of the parts? Studies of examiners' grading of individual papers and candidates' whole A-level examination performances. *Educational Studies*, *28*(2), 143–162. https://doi.org/10.1080/03055690220124588

Barham, P. J. (2012). An analysis of the changes in ability and knowledge of students taking A-level physics and mathematics over a 35 year period. *Physics Education*, *47*(2), 162. https://doi.org/10.1088/0031-9120/47/2/162

Barrance, R., & Elwood, J. (2018). National assessment policy reform 14–16 and its consequences for young people: Student views and experiences of GCSE reform in Northern Ireland and Wales. *Assessment in Education: Principles, Policy & Practice*, *25*(3), 252–271. https://doi.org/10.1080/0969594X.2017.1410465

Beaufort Research. (2022). *Survey of public opinions of non-degree qualifications in Wales 2022* (Research Report No. B02210-4). Qualifications Wales. https://qualificationswales.org/media/vjnesbdk/survey-of-public-opinions-of-non-degree-qualifications-in-wales-2022.pdf

Benton, T., & Bramley, T. (2015). *The use of evidence in setting and maintaining standards in GCSEs and A levels* (Assessment Research and Development). Cambridge Assessment. https://www.cambridgeassessment.org.uk/Images/459318-the-use-of-evidence-in-setting-and-maintaining-standards-in-gcses-and-a-levels.pdf

Black, H., Hall, J., Martin, S., & Yates, J. (1989). *The Quality of assessments: Case-studies in the National Certificate*. Scottish Council for Research in Education.

Blatchford, R. (2020). *The Forgotten Third: Do one third have to fail for two thirds to succeed?* John Catt Educational.

Boake, C. (2002). From the Binet–Simon to the Wechsler–Bellevue: Tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology*, *24*(3), 383–405. https://doi.org/10.1076/jcen.24.3.383.981

Brennan, R. L. (Ed.). (2006). *Educational Measurement* (4th edition). Rowman & Littlefield / Amer Council Ed Ac1 (Pre Acq).

Burdett, N., Houghton, E., Sargent, C., & Tisi, J. (2013). *Maintaining qualification and assessment standards: Summary of international practice*. Slough: National Foundation for Educational Research. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/605860/0113_NewmanBurdett_Maintaining_qualification_and_assessment_standards_V4_FINAL.pdf

Burge, B., & Benson, L. (2022). *National Reference Test results digest 2022.* National Foundation for Educational Research. Ofqual. https://www.gov.uk/government/publications/the-national-reference-test-in-2022/national-reference-test-results-digest-2022

Burgess, S. and Thomson, D. (2019). *Making the Grade. The impact of GCSE reforms on the attainment gap betweeen disadvantaged pupils and their peers*. The Sutton Trust. https://www.suttontrust.com/wp-content/uploads/2019/12/MakingtheGrade2019.pdf

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

Christie, T., & Forrest, G. M. (1981). *Defining Public Examination Standards*. Nelson Thornes.

Cizek, G., & Bunch, M. (2007). *Standard Setting*. SAGE Publications. https://doi.org/10.4135/9781412985918

Clarke, J. (1996). Modularising A levels in the Social Sciences: Some conclusions. *Social Science Teacher*, *26*(1), 33–35.

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. (1988). Combining grades from different assessments: How reliable is the result? *Educational Review*, *40*(3), 361–382. https://doi.org/10.1080/0013191880400308

Cresswell, M. J. (1994). Aggregation and awarding methods for National Curriculum assessments in England and Wales: A comparison of approaches proposed for Key Stages 3 and 4. *Assessment in Education: Principles, Policy & Practice*, *1*(1), 45–62. https://doi.org/10.1080/0969594940010104

Cresswell, M. J. (1996). Defining, setting and maintaining standards in curriculum-embedded examinations: Judgemental and statistical approaches. In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, Developments, and Statistical Issues: A Volume of Expert Contributions* (pp. 57–84). Wiley.

Cresswell, M. J. (2000). The role of public examinations in defining and monitoring standards. In H. Goldstein & A. F. Heath (Eds.), *Educational Standards* (pp. 69–109). Oxford University Press.

Cresswell, M. J. (2003). *Heaps, Prototypes and Ethics: The Consequences of Using Judgements of Student Performance to Set Examination Standards in a Time of Change*. Institute of Education, University of London.

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

Cuff, B. M. P. (2017). *An exploratory investigation into how moderators of non-examined assessments make their judgements* (Ofqual/17/6252). Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/633231/Ofqual-17-6252-Moderation_Report__Cuff__2017_.pdf

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.

Cumming, J. J. (2020). Senior secondary school assessment and standard-setting in Queensland, Australia: Social context and paradigmatic change. *Assessment in Education: Principles, Policy & Practice*, *27*(2), 160–177. https://doi.org/10.1080/0969594X.2019.1684877

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

El Masri, Y. H., Ferrara, S., Foltz, P. W., & Baird, J.-A. (2017). Predicting item difficulty of science national curriculum tests: The case of key stage 2 assessments. *The Curriculum Journal*, *28*(1), 59–82. https://doi.org/10.1080/09585176.2016.1232201

Fairbairn, D. J. (1988). Chapter 3: Pupil profiling: New approaches to recording and reporting achievement. In R. Murphy, H. Torrance, D. J. Fairbairn, D.

Pennycuick, & H. G. Macintosh, *The Changing Face of Educational Assessment*. Open University Press.
Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*, 171–191. https://doi.org/10.1037/0033-2909.101.2.171
Forster, M. (2011). The challenges of ensuring year-on-year comparability when moving from linear to unitised schemes at GCSE. *Research Matters: A Cambridge Assessment Publication*, *2*. https://www.cambridgeassessment.org.uk/Images/567586-the-challenges-for-ensuring-year-on-year-comparability-when-moving-from-linear-to-unitised-schemes-at-gcse.pdf
Furr, R. M. (2021). *Psychometrics: An Introduction* (4th edition). Sage Publications.
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
Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, *18*, 519–521.
Glaser, R., & Nitko, A. (1971). Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd edition). American Council on Education.
Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal*, *5*(2), 211–220.
Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, *33*(2), 234–246. https://psycnet.apa.org/doi/10.1111/j.2044-8317.1980.tb00610.x
Good, F. J., & Cresswell, M. J. (1988). *Grading the GCSE*. Secondary Examinations Council.
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*. https://www.cambridgeassessment.org.uk/Images/507975-a-review-of-research-about-writing-and-using-grade-descriptors-in-gcses-and-a-levels.pdf
Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, *24*(4), 355–366. https://doi.org/10.1177/01466210022031804
Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, *48*(1), 1–47. https://doi.org/10.3102/00346543048001001
Hart, J., Howieson, C., & Semple, S. (2010). *Recognising achievement literature review and model for managing recognition processes*. Education Analytical Services, Scottish Government.

Heinrich, M., & Stringer, N. (2012). *The effects on schools and pupils of modularising GCSEs*. Centre for Education Research and Practice.

Hipkiss, A., Woods, K. A., & McCaldin, T. (2021). Students' use of GCSE access arrangements. *British Journal of Special Education*, *48*(1), 50–69. https://doi.org/10.1111/1467-8578.12347

Holmes, S., Khan, A., Zanini, N., & Black, B. (2020). *Predicting predictability: Investigating question paper predictability and the factors that influence this through a question prediction exercise* (Research and Analysis). Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/940330/6714_1_predicting_predictability_fullreport.pdf

Howard, E., & Black, B. (2017). *Evaluation of reviews of marking and moderation 2016* (Ofqual/17/6253; Study and Survey). Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/633232/Ofqual-17-6253-ROMM_-Evaluation__Howard__Black__2017_.pdf

Hutchinson, C., & Hayward, L. (2005). The journey so far: Assessment for learning in Scotland. *The Curriculum Journal*, *16*(2), 225–248. https://doi.org/10.1080/09585170500136184

Independent Commission on Examination Malpractice. (2019). *Report of the Independent Commission on Examination Malpractice*. Joint Council for Qualifications. https://www.jcq.org.uk/examination-system/imc-home/

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. Gray, *Examination Standards: How Measures and Meanings Differ Around the World*. UCL IOE Press.

Isaacs, T., Zara, C., Herbert, G., & Coombs, S. (2013). *Key Concepts in Educational Assessment*. https://doi.org/10.4135/9781473915077

Jessup, G. (1989a). Foreword. In J. Burke (Ed.), *Competency Based Education And Training* (pp. ix–x). Routledge.

Jessup, G. (1989b). The emerging model of vocational education and training. In J. Burke (Ed.), *Competency Based Education And Training*. Routledge.

King, M. R. & chatGPT. (2023). A conversation on artificial intelligence, chatbots, and plagiarism in higher education. *Cellular and Molecular Bioengineering*, *16*(1), 1–2. https://doi.org/10.1007/s12195-022-00754-8

Klenowski, V., & Wyatt-Smith, C. (2010). Standards, teacher judgement and moderation in contexts of national curriculum and assessment reform. *Assessment Matters*, *2*, 107–131. https://doi.org/10.18296/am.0078

Klenowski, V., & Wyatt-Smith, C. (2014). *Assessment for Education: Standards, Judgement and Moderation*. https://doi.org/10.4135/9781526401878

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. https://doi.org/10.1371/journal.pone.0222279

Linn, R. (1995). *Assessment-Based Reform: Challenges to Educational Measurement* [Lecture]. William H. Angoff Memorial Lecture Series, ETS Policy Information Center.

https://www.ets.org/research/policy_research_reports/publications/publication/1995/buoa.html
Linn, R. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4–16. https://doi.org/10.3102/0013189X029002004
Linn, R., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of claims that 'Everyone is above average'. *Educational Measurement: Issues and Practice*, *9*(3), 5–14. https://doi.org/10.1111/j.1745-3992.1990.tb00372.x
Markowitsch, J., & Luomi-Messerer, K. (2008). Development and interpretation of descriptors of the European Qualifications Framework. *European Journal of Vocational Training*, *THEMATIC ISSUE: THE EUROPEAN QUALIFICATIONS FRAMEWORK*. https://www.semanticscholar.org/paper/Development-and-interpretation-of-descriptors-of-Markowitsch-Luomi-Messerer/255fbd74dd939c153349981b0e38540b4d53d491
McGhee, E., & Masterson, J. (2022). Access arrangements for secondary students. Are they fit for purpose? *Support for Learning*, *37*(2), 244–262. https://doi.org/10.1111/1467-9604.12407
Meadows, M., & Black, B. (2018). Teachers' experience of and attitudes toward activities to maximise qualification results in England. *Oxford Review of Education*, *44*(5), 563–580. https://doi.org/10.1080/03054985.2018.1500355
Newton, P.E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 149–170. https://doi.org/10.1080/09695940701478321
Newton, P.E. (2011). A level pass rates and the enduring myth of norm-referencing. *Research Matters: A Cambridge Assessment Publication*, *2*, 7. https://www.cambridgeassessment.org.uk/Images/567580-a-level-pass-rates-and-the-enduring-myth-of-norm-referencing.pdf
Newton, P.E. (2019). Macro- and micro-validation: Beyond the 'five sources' framework for classifying validation evidence and analysis. *Practical Assessment, Research, and Evaluation*, *21*(1). https://doi.org/10.7275/f75k-1y75
Newton, P.E. (2020). *What is the Sawtooth Effect?* (Research and Analysis). Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/936339/What_is_the_Sawtooth_Effect.pdf
Newton, P.E. (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., Baird, J., Goldstein, H., Patrick, H. & Tymms, P. (Editors) (2007) *Comparability of UK public examinations*. QCA book publication.
Newton, P. M. (2018). How common is commercial contract cheating in higher education and is it increasing? A systematic review. *Frontiers in Education*, *3*. https://www.frontiersin.org/articles/10.3389/feduc.2018.00067
Noyes, A., & Sealey, P. (2011). Managing learning trajectories: The case of 14–19 mathematics. *Educational Review*, *63*(2), 179–193. https://doi.org/10.1080/00131911.2010.534768
Office for National Statistics. (2022). *Population and household estimates, England and Wales—Office for National Statistics*. Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigr

ation/populationestimates/bulletins/populationandhouseholdestimatesenglandandwales/census2021unroundeddata
Ofqual. (2015a). *A comparison of expected difficulty, actual difficulty and assessment of problem solving across GCSE Maths Sample Assessment Materials*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/605887/2015-05-21-gcse-maths-research-on-sample-assessment-materials.pdf
Ofqual. (2015b). *Research into alternative marking review processes for exams* (Ofqual/15/5804). Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/605576/Research_into_alternative_marking_review_processes_for_exams.pdf
Opposs, D. (2016). Whatever happened to school-based assessment in England's GCSEs and A levels? *Perspectives in Education*, *34*(4), Article 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. Gray, *Examination Standards: How Measures and Meanings Differ Around the World*. UCL IOE Press.
Orr, L., & Forrest, G. M. (1984). *Investigation into the relationship between grades and assessment objectives in History and English examinations.* (Joint Research into Criterion Referencing of Grades at 16+). Joint Matriculation Board and Schools Council.
Pearson. (2017). Pearson Edexcel Level 2 NVQ Diploma in Housekeeping. Pearson Education Limited. https://qualifications.pearson.com/en/qualifications/nvq-and-competence-based-qualifications/hospitality-travel-tourism-events/housekeeping-l2.html
Philip, H. L. (2009). *The Higher Tradition, Volume 2: A history of public examinations in Scottish schools and how they influenced the development of secondary education.* https://era.ed.ac.uk/handle/1842/9429
Pinot de Moria, A. (2020a). *Common centres: In the context of maintenance of standards for the GCSE*. Qualifications Wales. https://qw-website-prod-master.azurewebsites.net/media/ji5njiko/common-centres.pdf https://qw-website-prod-master.azurewebsites.net/media/hpuecush/canolfannau-cyffredin.pdf
Pinot de Moira, A. (2020b) *The impact of coursework on attainment dependent on student characteristics. A study based on GCSE and A level outcomes between 2004 and 2017*. Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/896472/The_impact_of_coursework_on_attainment_dependent_on_student_characteristics.pdf
Pinot de Moira, A., Meadows, M., & Baird, J. (2020). The SES equity gap and the reform from modular to linear GCSE mathematics. *British Educational Research Journal*, *46*(2), 421–436. https://doi.org/10.1002/berj.3585
Pollitt, A., Ahmed, A., Baird, J.-A., Tognolini, J., & Davidson, M. (2008). *Improving the quality of GCSE assessment*. Qualifications and Curriculum Authority. https://www.researchgate.net/profile/Jim-Tognolini/publication/228735810_Improving_the_Quality_of_GCSE_Assessme

nt/links/02e7e526dd29d9ef78000000/Improving-the-Quality-of-GCSE-Assessment.pdf
Pollitt, A., Hughes, S., Ahmed, A., Fisher-Hoch, H., & Bramley, T. (1998). T*he effects of structure on the demands in GCSE and A level questions*. Qualifications and Curriculum Authority. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/605189/0198_AlastairPollitt_et_al_GCSE-ALevel-question-demands.pdf
Popham, W. J. (1992). A tale of two test-specification strategies. *Educational Measurement: Issues and Practice*, *11*(2), 16–22. https://doi.org/10.1111/j.1745-3992.1992.tb00235.x
Popham, W. J. (1993). Educational testing in America: What's right, what's wrong? A criterion-referenced perspective. *Educational Measurement: Issues and Practice*, *12*(1), 11–14. https://doi.org/10.1111/j.1745-3992.1993.tb00517.x
Popham, W. J. (1994). The instructional consequences of criterion referenced clarity. *Educational Measurement: Issues and Practice*, *13*(4), 15–18. https://doi.org/10.1111/j.1745-3992.1994.tb00565.x
Popham, W. J. (2014). Criterion-referenced measurement: Half a century wasted? https://www.ascd.org/el/articles/criterion-referenced-measurement-half-a-century-wasted
Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement*, *6*(1), 1–9. https://doi.org/10.1111/j.1745-3984.1969.tb00654.x
QCAA. (2019). English General Senior Syllabus 2019: Assessment. Queensland Curriculum and Assessment Authority. https://www.qcaa.qld.edu.au/senior/senior-subjects/english/english/assessment
QCAA. (2023a). 9.5 Endorsement (Units 3 and 4). Queensland Curriculum and Assessment Authority. https://www.qcaa.qld.edu.au/senior/certificates-and-qualifications/qce-qcia-handbook/9-internal-assessment-qa/9.5-endorsement
QCAA. (2023b). 1.3 Defining Characteristics of the QCE System. Queensland Curriculum and Assessment Authority. https://www.qcaa.qld.edu.au/senior/certificates-and-qualifications/qce-qcia-handbook/1-senior-schooling-qld/1.3-defining-characteristics
QCAA. (2023c). 7.3 Assessment Requirements. Queensland Curriculum and Assessment Authority. https://www.qcaa.qld.edu.au/senior/certificates-and-qualifications/qce-qcia-handbook/7-the-assessment-system/7.3-assessment-requirements
QCAA. (2023d). 9.6 Confirmation (Units 3 and 4). Queensland Curriculum and Assessment Authority. https://www.qcaa.qld.edu.au/senior/certificates-and-qualifications/qce-qcia-handbook/9-internal-assessment-qa/9.6-confirmation-units-3-4
Qualifications Wales. (2019). Criteria for Recognition to Award GCSE/GCE Qualifications [Regulatory Document]. Qualifications Wales. https://qualifications.wales/media/ayxb1dwz/criteria-for-recognition-to-award-gcse-gce-qualifications.pdf

Qualifications Wales. (2021a). *Qualified for the future: The right choice for Wales | Our decisions*. Qualifications Wales. https://qualificationswales.org/media/iehjpxeb/qualified-for-the-future-our-decisions.pdf

Qualifications Wales. (2021b). Standard Conditions of Recognition [Regulatory Document]. Qualifications Wales. https://www.qualificationswales.org/media/p54jkkfe/standard-conditions-of-recognition.pdf

Qualifications Wales. (2022a). *Malpractice in GCSE, AS and A level in Wales: Summer 2022 exam series*. Qualifications Wales. https://qualifications.wales/publications-resources/malpractice-in-gcse-as-and-a-level-in-wales-summer-2022-exam-series/

Qualifications Wales. (2022b). *Provisional entries for summer 2022*. https://qualifications.wales/publications-resources/provisional-entries-for-summer-2022/

Qualifications Wales. (2023a). *GCSEs, AS & A levels*. Qualifications Wales. https://qualifications.wales/information-support/qualifications-available-in-wales/gcses-as-a-levels/

Qualifications Wales. (2023b). *Made-for-Wales GCSEs: Main consultation report*. Qualifications Wales. https://qualifications.wales/media/dkcisr1u/made-for-wales-gcses-main-consultation-report.pdf

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

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

Shepard, L.A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice*, *9*(3), 15–22. https://doi.org/10.1111/j.1745-3992.1990.tb00374.x

Shepard, L.A. (1997). *Measuring Achievement: What does it mean to test for robust understanding?* William H. Angoff Memorial Lecture, Education Testing Service. https://www.ets.org/Media/Research/pdf/PICANG3.pdf

Smaill, E. (2020). Using involvement in moderation to strengthen teachers' assessment for learning capability, *Assessment in Education: Principles, Policy & Practice*, 27:5, 522-543, DOI: 10.1080/0969594X.2020.1777087

SQA. (n.d.). *Understanding standards*. https://www.understandingstandards.org.uk/

SQA. (2016). *National course design and assessment. SQA fieldwork visits – engagement and focus group discussions with centres delivering new national qualifications*. Scottish Qualifications Authority. https://www.sqa.org.uk/sqa/files_ccc/SQA_Fieldwork_visits.pdf

SQA. (2017). National 4 Applications of Mathematics Course Specification. Scottish Qualifications Authority. https://www.sqa.org.uk/files/nq/AppsofMathsCourseSpecN4.pdf

SQA. (2019a). Guidance on the Implementation of Graded Units in Higher National Certificates and Diplomas (No. CA7952). Scottish Qualifications Authority.

SQA. (2019b). Guide to Assessment. Scottish Qualifications Authority.

STA. (2018). Teacher Assessment Frameworks at the End of Key Stage 2 (STA/19/8302/e). Standards and Testing Agency. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1119094/2018-19_teacher_assessment_frameworks_at_the_end_of_key_stage_2.pdf

StatsWales. (2022). Schools by Local Authority, Region and Type of School. Welsh Government. https://statswales.gov.wales/Catalogue/Education-and-Skills/Schools-and-Teachers/Schools-Census/Pupil-Level-Annual-School-Census/Schools/schools-by-localauthorityregion-type

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

Stringer, N. (2014). *The achieved weightings of assessment objectives as a source of validity evidence* (Ofqual/14/5375). Ofqual.

Taverner, S., & Wright, M. (1997). Why go modular? A review of modular A-level Mathematics. *Educational Research*, *39*(1), 104–112. https://doi.org/10.1080/0013188970390108

Taylor, R. (2016). The effects of accountability measures in English secondary schools: Early and multiple entry to GCSE Mathematics assessments. *Oxford Review of Education*, *42*(6), 629–645. https://doi.org/10.1080/03054985.2016.1197829

Taylor, R., & Opposs, D. (2018). Standard setting in England. In J.-A. Baird, T. Isaacs, D. Opposs, & L. Gray (Eds.), *Examination Standards: How Measures and Meanings Differ Around the World*. UCL IOE Press.

The Newsroom. (2011, January 14). New Curriculum for Excellence 'will harm teaching of history'. *The Scotsman*. https://www.scotsman.com/news/new-curriculum-excellence-will-harm-teaching-history-1688982

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

Trahan, L. H., Stuebing, K. K., Fletcher, J. M., & Hiscock, M. (2014). The Flynn effect: A meta-analysis. *Psychological Bulletin*, *140*(5), 1332–1360. https://doi.org/10.1037/a0037173

Tuxworth, E. (1989). Competence based education and training: Background and origins. In J. Burke (Ed.), *Competency Based Education And Training*. Routledge.

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review*, *32*, 100374. https://doi.org/10.1016/j.edurev.2020.100374

Vidal Rodeiro, C. L., & Nádas, R. (2012). Effects of modularity, certification session and re-sits on examination performance. *Assessment in Education: Principles, Policy & Practice*, *19*(4), 411–430. https://doi.org/10.1080/0969594X.2011.614218

Wei Lee, M., & Newton, P. N. (2021) *Systematic divergence between teacher and test-based assessment: Literature review*. Ofqual. https://www.gov.uk/government/publications/systematic-divergence-between-teacher-and-test-based-assessment/systematic-divergence-between-teacher-and-test-based-assessment-literature-review

Welsh Government. (2014). *Subject principles for GCSE Welsh Language* (No. 151/2014). Welsh Government. https://dera.ioe.ac.uk/20941/1/140829-welsh-language-2015-en.pdf

Welsh Government. (2022). *School improvement guidance: Framework for evaluation, improvement and accountability – Hwb* (Evaluation, Improvement and Accountability). Welsh Government. https://hwb.gov.wales/evaluation-improvement-and-accountability/school-improvement-guidance-framework-for-evaluation-improvement-and-accountability

Wiliam, D. (1996). Standards in examinations: A matter of trust? *The Curriculum Journal*, *7*(3), 293–306. https://doi.org/10.1080/0958517960070303

Wilmut, J. and Rose, J. 1989. *The Modular TVEI Scheme in Somerset: its concept, delivery and administration*, London: Report to the Training Agency of the Department of Employment.

WJEC. (1988). English and English Literature GCSE and GCE Syllabus 1988. Welsh Joint Education Committee.

WJEC. (2022). Guide to Resit Requirements. Welsh Joint Education Committee. https://www.wjec.co.uk/media/fvod4qdz/guide-to-resits-2022.pdf?language_id=1

Wolf, A. (1995). *Competence-Based Assessment*. (Assessing Assessment). Open University Press.

Wolf, A. (1996). Vocational assessment. In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, Developments, and Statistical Issues: A volume of expert contributions* (pp. 209–230). Wiley.

Wolf, A. (2001). Competence-based assessment. In J. Raven & J. Stephenson, *Competence in the Learning Society* (pp. 453–466). Peter Lang.

9 Appendix A: Advisory Group remit and membership

Remit

1. Oxford University's Department of Education has received funding from Qualifications Wales to conduct a two-year research project to consider the meaning and communication of standards in Welsh GCSEs (and A-levels) now, and under alternative definitions.
2. The purposes of the project are to
   - consider how standards are embedded in GCSEs and A-levels,
   - relate this to current definitions of standards,
   - consider how best to communicate this to stakeholders,
   - develop teaching materials for use with teachers and to
   - design a study to investigate the implications of criterion-referencing in the context of Welsh GCSEs.
3. The Oxford Advisory Group for the Standards in Welsh GCSEs project has been established to provide independent advice to the researchers throughout the project.

Membership

4. The Oxford Advisory Group shall consist of between 9 and 12 members in total. The members are appointed from the education, education policy and assessment communities to bring expertise and experience, including international perspectives.
5. Appointments are made by Oxford for the duration of the project.

Purpose and role

6. Members of the Advisory Group are asked to maintain an overview of the work conducted by the project, consider issues and provide expert advice in relation to research and analysis conducted as part of this project as well as the dissemination of research findings.
7. Members of the Advisory Group may also be asked to act as participants in the research. For example, they may be asked to complete short surveys or take part in interviews to inform thinking with regard to particular issues.
8. To assist with the veracity of findings and with the impact of the research, the Advisory Group may also advise on key organisations, events or individuals that the project should interact with.

Meetings

9. The Advisory Group will be chaired by a member of the Oxford Project Team.
10. The Advisory Group shall meet approximately 3 times.

11. Members have been carefully selected for their expertise and experience. For this reason, attendance should not be covered by alternates without a clear rationale that has been agreed with the Project Team.

12. The Chair may invite other individuals to attend meetings, for example to hear a particular stakeholder's or expert's point of view on a matter. The invitation may be for the entire meeting or for one or more specific item(s).

Reporting

13. Oxford will be responsible for organising and taking notes of the Advisory Group meetings. Notes will be used by the Project Team and for future meetings as a reference.

Membership

Siôn Amlyn	Policy and Casework Official at NASUWT Cymru Teachers' Union
Jonathan Angell	Headteacher and Director of Eastern Community Campus, Cardiff
Neil Butler	National Official for Wales, NASUWT
Richard Daugherty	Emeritus Professor at Aberystwyth University, Honorary Professor at Cardiff School of Social Sciences
Geoff Evans	Headteacher at Ysgol Y Strade, Llanelli
Georgina Haarhoff	Deputy Director for Curriculum, Assessment and School Improvement in the Education Department at the Welsh Government
Richard Harry	Director of Qualifications and Assessment at WJEC
Eithne Hughes	ASCL Cymru Director
Jamie Insole	Policy and Political Official for the University and College Union in Wales
Alison Matthews	Deputy Director of Undergraduate Admissions at the University of Oxford
Claire Morgan	Strategic Director at Estyn
Mary van den Heuvel	Senior Policy Officer at NEU Cymru

10 Appendix B : Organisational responsibilities for qualification standards

Assessment in Wales can be traced back to the introduction of public exams in the late 19th and early 20th centuries. Over time, the Welsh assessment system became increasingly aligned with the English system, with GCSEs and A-levels being introduced and developed in a similar manner in both countries. However, in recent years, there has been a growing recognition of the need for the Welsh assessment system to be more closely aligned with the Welsh education system and its priorities. This has resulted in a move towards a Wales-specific market for GCSEs and A-levels, with qualifications being developed specifically for the Welsh market. Several organisations have complementary roles in the general qualifications system in Wales (Table 5).

Table 5 Institutional oversight of qualification standards in Wales

Welsh Government	Responsible for the development and implementation of education policy in Wales. This includes the ownership of the programmes of study for each curriculum subject and setting the overarching framework for qualifications and assessments. Government is also responsible for school improvement and evaluation arrangements, which can have significant effects on qualification design and delivery.
Qualifications Wales	An independent body that regulates the development and delivery of qualifications in Wales, including those that are government funded. As such, it operationalises government policy through the creation of regulations which exam boards must follow. In particular, it is responsible for overseeing the setting of standards for GCSEs and A-levels, ensuring the quality and comparability of qualifications, and advising the Welsh Government on qualifications policy.
WJEC	An exam board that provides a range of qualifications in Wales. Within the regulations set by Qualifications Wales, it designs and delivers GCSEs and A-levels. This includes publishing specifications, writing and delivering exam papers, marking them, setting coursework and managing the moderation of teacher marks, setting grade boundaries and managing marking reviews and appeals.
Estyn	The education inspectorate for Wales. It is responsible for inspecting and reporting on the quality of education in Wales, including the teaching and learning of qualifications such as GCSEs. The work of Estyn therefore impacts the standard of student performance.
UCAS	A UK-wide organisation responsible for processing applications to higher education institutions in the UK. It allocates points to qualification outcomes through the UCAS Tariff. It also provides information and guidance to students and parents on qualifications, including GCSEs and A-levels.

11 Glossary of terms

Word	Definition
Achievement	The grade or level that a learner has been able to reach.
Assessment	(verb) The act of judging or deciding the amount, value, quality or importance of something, or the judgment or decision that is made. (noun) A task designed to elicit evidence of specific knowledge and/or skills.
Attainment	The standard of performance that a learner has been able to produce in given conditions.
Attainment-referencing	Candidates receive grades that reflect their holistic attainment in the qualification at a standard which is comparable with the attainment required for that outcome in previous years' qualifications.
Cohort-referencing	Candidates receive grades that tell us where they rank in relation to the cohort who took the qualification in the same year.
Comparability	The degree to which it is possible to compare the standards of an assessment over time or between similar qualifications.
Comparable outcomes	Candidates receive, as a group, comparable grade outcomes to those which they would have received had they followed the course before a reform and taken the old qualification.
Controlled assessment	A GCSE assessment conducted by teachers within schools and colleges under varying levels of control to mitigate the risk of malpractice.
Coursework	Assessment conducted by teachers within schools and colleges.
Criterion-referencing	Candidates receive grades that tell us whether they met predetermined performance criteria.
Examination	A test conducted under controlled conditions and with a pre-specified time limit.
Examination centres	Approved centres where candidates can sit examinations or other tests. Most schools and colleges can be examination centres, but others exist for private candidates or for administration of non-school-based qualifications.
Examiners/Senior Examiners	Subject experts, often teachers or ex-teachers, responsible for setting assessments and mark schemes, moderation, marking and standard setting.

External assessment	Assessment that is external to the school, usually conducted by the examination board and typically examinations.
General qualification	Qualifications that are not linked to particular fields of work or employment, but instead assess a more generalised set of skills or capabilities, often linked to a particular subject area.
Grade boundary	The minimum mark needed to achieve a specified grade. Sometimes called a cut-score.
Grade inflation	The observed effect in which the proportion of individuals achieving a particular grade or better increases year on year. Grade inflation is often referenced as evidence of a lowering of qualification standards, although this is not universally accepted.
Grading	The act of translating raw marks or descriptors into grades, usually in the form of ranked letters (e.g. A*-F) or pass/fail.
Internal assessment	Teacher assessment that is internal to the school, typically coursework. Often referred to as 'non-examined assessment'.
Linear examination	Students sit all of their exams in one series at the end of the course of study.
Marking	The process of assigning marks (scores) to answers to questions or tasks.
Moderation	The process of ensuring that grades or marks are awarded consistently within and between schools or centres.
Modular qualification	The totality of the assessment is broken into discrete units for assessment, the results of which are combined to give an overall qualification outcome.
Norm-referencing	Candidates receive grades that tell us where they rank in relation to the population of students who could have taken the qualification in any year.
Objective test	Multiple choice test, which is usually machine-marked.
Portfolio	A range of different sources of evidence gathered by an individual learner to demonstrate proficiency in a certain topic, subject or discipline.
Predictive validity	The extent to which predictions made based on measurements/assessments come true.
Professional qualification	Specific qualification required to work in a certain field. Occasionally referred to as an 'advanced vocational qualification' or a 'licence to practice qualification'.
Qualification	Officially certified confirmation of the level of proficiency in a specified area.

Reasonable adjustments	Adjustments made to an assessment to enable disabled learners to demonstrate their knowledge, skills and understanding.
Regulator	An officially recognised body responsible for regulating or supervising a particular industry. In the case of educational assessment in Wales, Qualification Wales is the regulator for all recognised qualifications other than university degrees.
Reliability	The extent to which scores are consistent.
Sawtooth Effect	The observed effect in which cohort performance on high-stakes assessments drops after assessment reform and then improves over time as assessment familiarity increases.
Special consideration	Consideration given to learners who have temporarily experienced illness or injury, or some other event outside of their control, which has impacted their ability to take an assessment or demonstrate their attainment in an assessment.
Standard	A standard is a pre-agreed reference against which student outcomes can be evaluated.
Standard setting	In the context of GCSE, it is the process of transforming marks into grades.
Terminal assessment	Assessment at the end of a course of study.
Validity	The extent to which the measurement/assessment measures/assesses that which it claims to measure/assess.
Viva	An interview in which examiners question an examinee to determine their proficiency in a certain area. Often vivas will revolve around a pre-submitted piece/portfolio of work.
Vocational qualification	Qualification linked to a particular field of work, or vocation. They often involve a practical element and the process of standardsetting may be determined by the relevant industry or professional body.