Centre for Education Research and Practice

Developing grade descriptions for the new GCSEs: Considerations and challenges

Stuart Cadwallader

Abstract

Grade descriptions are to be developed for the new GCSE qualifications that will be first taught from September 2015 (Ofqual, 2014). This paper discusses the potential role of grade descriptions in the new qualifications by exploring their purpose and the issues which impact on their effectiveness. It is argued that grade descriptions, in their current form, are not useful for either of their two primary purposes: supporting teachers as they prepare students for assessment and supporting examiners as they set grade boundaries during awarding. Possible methods for developing new descriptions are outlined, with particular attention given to an empirical approach whereby data from live examinations is used to identify items which discriminate between candidates at different grades (e.g. Greatorex, Johnson, & Frame, 2001). It is proposed that the current approach to grade descriptions should be discontinued and replaced with post-hoc exemplification of grades that will better meet the needs of teachers while also more accurately reflecting the nature of the assessment system.

Introduction

GCSEs and A-levels, the major national qualifications in England, are currently being reformed and the new specifications are to be first taught from September 2015 (Gove, 2013a, 2013b). For the new GCSEs, policy makers have stipulated that assessment is to demand higher performance for the award of key grades: 'At the level of what is widely considered to be a pass (currently indicated by a grade C), there must be an increase in demand, to reflect that of high-performing jurisdictions' (Gove, 2013b, p. 2). In response, stakeholders have expressed a desire to understand how the new grade scale will compare to the old one (YouGov, 2013). Teachers consider it important that they have a clear understanding of what they are aiming for so that they are best placed to help their students to reach these higher levels of performance. Clarity is also sought by end users of the new GCSEs, those who will use the grades as the basis for selecting applicants for employment or education.

To assist with the transition, the Office of Qualifications and Examinations Regulation (Ofqual) are promoting a rethink about grade descriptions (sometimes referred to as grade descriptors or performance descriptions). Greatorex (2005) defines grade descriptions as '…indicators which exemplify the qualities candidates are likely to exhibit if they achieve a particular grade' (p.9). They are essentially a qualitative articulation of the skills and attributes associated with performance at that grade. Grade descriptions can be thought of as an important link between examiners and teachers because they serve to make examiners' implicit understanding of performance standards explicit (Greatorex, 2002; Sadler, 1987).

Ofqual have posited a plan for the first year of the reformed GCSEs:

> *We propose that for the first year we will write three illustrative and general grade descriptions for grades towards the top, the bottom and in the middle of the grade range[1].These descriptions will be primarily to help teachers understand the standard of performance expected at these points. So far as possible, these descriptions will describe the key characteristics and levels of achievement we might expect. We do not propose that these descriptions, which will be untested, should be used for awarding in the first year; we propose awarding should be based on statistical predictions. We will test these descriptions against the performance actually demonstrated in the early years and refine them as necessary. We will consider in due course whether in the future such descriptions could have any role in awarding.*
>
> (Ofqual, 2014, p. 20)

The purpose of the new grade descriptions is clear: they are to provide teachers with an understanding of the performance standards that will be required. Additionally, the fact that grade descriptions have the potential to play a role in awarding grades to candidates is acknowledged and the need for future evaluation and analysis is made clear. This paper sets out to engage with the literature about grade descriptions and discuss some of the issues that might be relevant for the new GCSEs, both at launch and in the future. A good starting point is to set the discussion in context by describing the origins of grade descriptions.

From criteria to descriptor: The story of grade descriptions in England

Grade descriptions for GCSE have their roots in ‘grade related criteria’, which were developed, but not implemented, when the qualification was first introduced in the 1980s. The rationale for introducing these criteria was that teachers and pupils should be able to set themselves subject-specific goals and would therefore require explicit instruction as to the necessary knowledge, understanding and skills expected at each grade (see Gipps, 1990 for a summary). Despite the apparent similarity to grade descriptions, ’grade related criteria’ had a significantly different purpose. They were to be tied to the introduction of a National curriculum in England, with the plan being that the new curriculum would be assessable via fixed criteria using tests taken by pupils in specific school years (for more information see Whetton, 2009). Essentially, if these criteria had been brought into effect, it would have meant that GCSEs would have been fully criterion referenced, with a candidate’s grade based on their ability to meet *all* of the relevant criteria. The criteria would have explicitly specified the performance standard by representing the qualities that a candidate’s work would *have* to exhibit for the relevant grade to be awarded.

After several iterative attempts by the working parties to develop grade related criteria it became apparent that a system of assessment governed by highly specified criteria was not viable for national general qualifications. There were a range of reasons, but the primary one was that it was deemed unfeasible to break the content of the subjects down into such small units (the criteria) while maintaining valid and fair assessment. It is worth illustrating this point. The emerging criteria were to be divided into *domains* and *sub-domains*, defined by ‘coherent and defined areas of knowledge, understanding and skills’ (Gipps, 1990, p. 83). For example, performance in English may have been divided into the domains of ‘writing’ and ‘reading’ and

[1] To help benchmark the new performance standard, Ofqual have collected and reviewed performance descriptors from countries whose students tend to perform well in international tests (Ofqual, 2014).

then 'writing' may have been divided again into the sub-domains of 'content', 'structure', and 'use of language'. However, Cresswell (1987a) describes how, in order to make the criteria sufficiently meaningful, it was necessary to divide content across a large number of domains, sub-domains and even sub-sub-domains. The criteria that were drafted were simply too numerous, complex and multi-faceted to be useful (Gipps, 1990).

Various methods for salvaging the grade related criteria were attempted and the developers switched their attention to describing desirable 'attributes' at each grade. The attributes described behaviour rather than specified content and were organised into 'performance matrices' (see Cresswell, 1987b). These performance matrices represented a less prescribed form of the criteria because there were multiple ways in which a candidate could demonstrate the various attributes. Ultimately, this final approach provided the basis for the more loosely defined grade descriptions that are in use for the current version of GCSEs.

The current system operates using what can be best described as 'weak criterion referencing' (Baird, Cresswell, & Newton, 2000). In a system of weak criterion referencing, candidates do not have to meet the performance standard against all of the criteria in the assessment to achieve a given grade; they are allowed to compensate for poor performance in one or more areas with strong performance in others (more on this later). Grade descriptions therefore offer only an *indication* of the level of skill and knowledge that is likely to be characteristic of a candidate who achieves a given grade (Greatorex, 2005).

Descriptions are currently provided for each judgemental grade[2] as part of the specification for each GCSE qualification. When presenting descriptions, the awarding bodies make the following four points explicit (e.g. AQA, 2012; OCR, 2013):

1. Grade descriptions are for general guidance only.
2. In practice, the grade awarded to an individual candidate will depend on their overall performance against the *assessment objectives* rather than how closely their performance matches the grade description.
3. Grade descriptions are to be interpreted in relation to the content of the specification; it is not intended that they define the content.
4. If a candidate does not perform well against one or more aspects of a grade description, this may be balanced by better performance on other aspects.

There is considerable variation between subjects with regard to the content of grade descriptions but **Appendix A** provides an example of how they are often structured. Arguably, the level of detail expressed in the current grade descriptions is rather limited – they do not provide sufficient detail for teachers to plot a course to a given grade, only a broad overview of how candidates at each grade might perform. An understanding of this historical context is useful for appreciating some of the key issues which will now be discussed.

Constraints imposed by the system of assessment and awarding

If grade descriptions are to be effective it is important that they are not undermined by other elements of the assessment system. There are two elements that limit the level to which descriptions can be precise and explicit and may therefore constrain the extent to which they can be useful for teachers and examiners. The first is the compensatory nature of grades, which is determined by the role that criteria play during awarding, and the second is the level of stability in performance standards, which is partially dictated by the approach that is taken to maintaining standards. Each of these two points shall now be discussed in turn.

[2] The judgemental grades are A, C, and F for GCSE.

Compensation across criteria in a system of weak criterion referencing

As has been discussed, a 'strongly' criterion referenced system has a series of explicit hurdles whereby a grade cannot be achieved unless *all* relevant criteria are met. This may be appropriate for certain types of qualification but in the context of general qualifications, such as the GCSE, this approach is arguably unfair because it heavily penalises those candidates who slip up against even a single criterion. For example, would it be fair for a candidate who is an outstanding creative writer (and meets the criteria for the top grade in this regard) to receive the lowest grade due to poor use of punctuation? This would be the case if criteria were not *compensatory*. As Cresswell (1986, p. 8) observed, a candidate's grade would be determined by the simplest task at which they had failed[3].

A compensatory system has clear advantages in terms of fairness to candidates and flexibility. However, there is a significant drawback to compensation: it undermines the purpose of grade description because it obscures the *meaning* of a grade. Successes and failures against the criteria can be combined in a multitude of ways to achieve a specific mark, meaning that two candidates could reach the same grade boundary but exhibit very different patterns of performance. As a result it becomes extremely challenging to describe or exemplify 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, 2001; Cresswell, 1987a).

A compensatory system is a necessity whenever the number of marks that a candidate has accrued across all assessment criteria are tallied and used as the basis for deciding their grade (Baird et al., 2000). According to the recent consultation the new qualifications are highly likely to retain the current approach to awarding grades and thus the 'weaker' form of criterion referencing which allows for compensation (Ofqual, 2014). Grade descriptions are therefore fundamentally limited with regard to how precise they can be – they can only describe performance in a very broad manner if they are to encapsulate the many ways in which candidates can reach a grade.

Stability of performance standards and the comparable outcomes approach

It is proposed that a comparable outcomes approach to maintaining standards will continue to be employed for the new qualifications (Ofqual, 2014). This also has an impact on what grade descriptions can realistically achieve. The comparable outcomes approach assumes that the proportion of candidates at a particular grade for a given subject will not vary between examination series if each cohort is similar in terms of prior attainment (Ofqual, 2013)[4]. The approach can broadly be described as cohort referencing, though the statistical evidence is balanced by the expert judgement of examiners, who review candidates' scripts around the judgemental grade boundaries to ensure that the statistical recommendations are plausible.

One of the main purposes of using statistical evidence in this way is to prevent cohorts who are taking exams in different series from being advantaged or disadvantaged due to variations in the demand of the assessment. A prime example would be when a new specification is introduced. Candidates taking the new specification are likely to underperform relative to those who took the established version because teachers, understandably, require time to adapt to the new content and assessment (see Jones, 2009 for a fuller explanation). It would be unfair

[3] The Diploma exhibited this problem. The qualification achieved a very low pass rate because the hurdles in its structure led to the heavy penalisation of poor performance against single criteria (see Isaacs, 2013).

[4] This is the general principle though there are acceptable reasons why outcomes may change between series (see Ofqual, 2013 for a thorough explanation).

for students to receive lower grades in the new specification than they would have achieved had they taken the previously established one. By assuming that, once the prior attainment of the cohort is accounted for, a similar proportion of candidates should achieve each grade, regardless of the series, this unfairness is nullified.

The problem is that an emphasis on comparable outcomes is not compatible with the use of grade descriptions. Essentially, the performance standard for a given grade can be different at each point in a specifications' life cycle, so it follows that the related grade description requires considerable flexibility. If a grade description was too specific it would be undermined by fluctuations in the performance standard that it was trying to describe. Unfortunately this reduces the usefulness of grade descriptions because the level of detail that they can provide is substantially stunted.

Supporting examiners when awarding the new GCSEs – Grade descriptions as the basis for awarding grades

The previous section has raised some significant issues with regard to how detailed and specific grade descriptions can be in the English assessment system. These issues notwithstanding, it is worth discussing how descriptions could be operationalised to best serve their proposed purposes for the new GCSEs. This section will focus on their use during awarding. Grade descriptions are currently available to GCSE awarding committees but it is unclear how frequently they are referred to in practice. Ofqual have decided not to provide them to support awarding for the new qualifications but have stated that it is worth considering whether they could be employed in the future (Ofqual, 2014). To explore this possibility, this section builds on the previous one to discuss how grade descriptions would need to be developed if they were to play an effective role. There are two issues; the first is technical and relates to where grade descriptions should be targeted, while the second is more overarching as it relates to the purpose of grade descriptions and how they could be used to *set* the initial performance standards for grading.

Describing minimum performance versus describing mid-grade performance

During awarding, examiners are seeking to identify the minimum mark at which a grade should be awarded to a candidate. However, the grade descriptions that are currently in use for GCSEs describe the typical performance of candidates who receive a given grade and are therefore targeted at the middle of that grade. This means that descriptions are currently not optimised for supporting the fine grained judgements that are required from examiners during awarding. If they were to instead target the minimum required performance they would be a more appropriate reference tool because they would be articulating the performance standard which the examiners were expected to maintain.

There are two linked problems here. Firstly, Greatorex (2003) suggests that it is easier to describe performance at mid-grade rather than at the minimum because there is a greater margin for error. It is possible for a description to 'miss' the middle of the grade by some way and yet still describe performance that is representative of that grade. Borderline performance is by its very nature a smaller target and therefore more difficult to identify and describe. Secondly, as discussed, the comparable outcomes approach allows the performance standard at a grade boundary to change from series to series. Such fluctuations in performance standard will have a greater impact at the boundary between grades than at mid-grade, where small deviations are less likely to impact on the validity of the grade description.

Aspirational approaches to developing grade descriptions

The method for deriving grade descriptions and their ultimate role in setting and maintaining standards is crucial to the extent to which they can serve examiners. Greatorex (2003) divides methodologies for developing descriptions in to two broad categories: *empirical* and *aspirational*. The grade descriptions that arise from these differing methodologies may look similar but are based on strikingly different underlying principles. Empirical approaches use data gained from examinations to provide post-hoc descriptions of the performance exhibited by candidates at different grades (more on this later). Aspirational approaches are ad-hoc, with grade descriptions developed to reflect the expectations of policy makers and assessors. Descriptions developed in this way are usually used as the basis for awarding grades to candidates and are therefore worth discussing in more detail.

If grade descriptions are to support awarding, a logical approach is to use them as the starting point for establishing the level of performance expected at each grade when a new qualification is first developed. If an aspirational approach is adopted, the standard of performance which policy makers want the qualification to encapsulate is articulated, with the help of assessment and education experts, and expressed in the grade descriptions. Subsequent grading is then carried out against these benchmarks and this standard is maintained throughout subsequent series. It is essentially a form of criterion referencing whereby the descriptions form the basis of initial grading decisions. This approach provides examiners with a foundation for their decisions during awarding (and teachers with a clear understanding of how grades are decided). This, as discussed earlier, was the initial plan for GCSEs (Cresswell, 1987a).

For example, in America the performance level description (PLD) has a clearly defined purpose in both the setting and maintenance of standards for the National Assessment of Educational Progress (NAEP). The creation of PLDs are a key step in setting the initial standard for the assessment and become instrumental when grade boundaries are set (Wyse, 2013). The development of PLDs and their use for the setting of cut scores clearly reflects their aspirational purpose. They are developed from the ground up, defined by the policy-makers' position regarding the performance or rigour that they desire for each level. The involvement of other stakeholders, such as educationalists, is incorporated to prevent the aspirations from being unrealistic or ungrounded in education theory (Perie, 2008). The final PLDs are often drafted on the basis of established education theory, such as Bloom's taxonomy (e.g. Anderson et al., 2014). Once the PLDs have been written, the first grade boundaries are then based on those who achieve this level, and this performance standard is later maintained using statistical tools such as test equating.

This has been a very brief and uncritical description of the American system but there are significant issues. For example, it is challenging to ensure that the initial aspirational PLDs remain the focus of standard maintenance once statistical information becomes a more powerful component in the process (Wyse, 2013). However, this approach clearly has desirable features and PLDs have a clearly defined role and purpose. Setting the standard in this way is undoubtedly helpful for examiners as it provides a clear and stable benchmark. Arguably, it is also helpful to policy makers, who are able to gauge whether the education system is becoming more or less effective at raising performance standards relative to a benchmark which they themselves have deemed desirable and appropriate[5].

[5] It should be noted that the instability of performance during the life cycle of a specification, an issue which the comparable outcomes approach is in place to mitigate against (see Jones, 2009), would somewhat limit the strength of this latter claim. Analysis of cohort performance across series would only become valid once a specification had been in use for long enough for it to reach a relatively stable state.

The question as to how an aspirational system would work if transplanted to assessment in England is not an easy one. The context of assessment in America is very different to that in England and a radical overhaul would be required to accommodate this type of approach. The main issue is that this type of ad-hoc aspirational system is essentially a form of criterion referencing and is therefore not compatible with the comparable outcomes approach. There are arguments for and against comparable outcomes (Baird et al., 2000) but it is fair to state that it would be problematic if the proportion of candidates at each grade was not monitored statistically and was therefore prone to significant fluctuation.

Ofqual (2014) have considered the risks in their consultation. For example, when New Zealand attempted to reform their assessment system and adopted a strongly criterion referenced approach there was a collapse in pass rates which was hugely problematic for pupils, schools and end users of the grades (Nash, 2005). The message in the case of New Zealand was that examiners could not tell in advance how well a cohort will do on un-trialled items against un-trialled criteria (Nash, 2005). If an aspirational approach were to be adopted for the new GCSEs in England there would have to be an uncomfortable acceptance that outcomes may be very low compared to previous versions of the qualification (at least initially) and that outcomes may change in unpredictable ways throughout the life cycle of each specification.

Conclusion: Grade descriptions should not be used to support awarding

Though aspirational approaches would provide grade descriptions that were central to assessment and therefore highly useful for examiners, it is clear that the approach is not compatible with the assessment system which will be in place for the new GCSEs. It is also clear that the comparable outcomes approach and the use of compensation in the assessment system combine to impose substantial limitations on how precise grade descriptions can be (as discussed in the previous section). These limitations would be further compounded if grade descriptions were to be targeted at the minimum requirement for a grade rather than at mid-grade, a step which would be necessary in order to make them useful for examiners.

Taking these issues together, it appears that grade descriptions are unsuited to the purpose of supporting examiners during awarding. Though they could be conceived as ‘supporting information’ to assist awarding committees as they decide on the plausibility of outcomes, as is currently the case, it could be argued that this would be tokenistic. Suggesting that they were appropriate for such a role would give them a veneer of precision and authority that they simply cannot achieve. The next question to consider is whether or not grade descriptions can be of value to teachers.

Supporting teachers in the first years of the new GCSEs and beyond

Ofqual (2014) have recognised the need to provide teachers with guidance about performance standards ahead of the first assessment series of the new GCSEs. Arguably, this is the point at which such support is required the most because teachers and learners are grappling with new qualification content and new forms of assessment. There are not previous examinations from which they can infer the standard that is likely to be necessary for a given grade, meaning that grade descriptions provide the only information that is available for reference.

A significant issue here is that grade descriptions are to be developed in an ad-hoc fashion. Teachers are to be provided with approximate descriptions of what the performance standards *may* be before the first assessment, but these descriptions will not be used in the process of actually *setting* performance standards. The actual performance standards will be set mainly on the basis of statistical evidence (comparable outcomes). It is therefore necessary for the initial grade descriptions to be imprecise estimates, meaning that they can only be presented to teachers on a highly tentative basis (Ofqual, 2014). Given the fundamental limitations to the extent to which grade descriptions can provide teachers with the detailed information that they are seeking, there is a danger that their importance may become exaggerated.

Grade descriptions can have unintended impacts on teaching and learning

It is worth considering whether grade descriptions are actually beneficial to teaching and learning. At their core, grade descriptions are qualitative statements and are therefore open to the interpretation of the reader. Even at their most precise, they encapsulate concepts that are often abstract. Two people reading a grade description, even one that is thoughtfully articulated, are likely to interpret the words slightly differently and therefore garner slightly differing understandings. Indeed, research by Richardson (2003a) suggests that it is even challenging to gain complete agreement between examiners on the same assessment regarding the exact structure and wording of descriptions.

Richardson (2003b) describes the process which was used to prepare grade descriptions for the World Class Test, a mathematics assessment that was targeted at high ability pupils between the ages of nine and thirteen. An empirical post-hoc methodology was employed, based on the qualitative scrutiny of candidates' scripts by teachers and principle examiners. Though the process itself ran relatively smoothly, a subsequent validation study, which drew on the expertise of test administrators and developers, suggested that the grade descriptions that had been produced had significant shortcomings:

> *…test developers were unanimous in their rejection of the proposed descriptors citing a range of reasons, including details of the text, construction methodology, ownership and standards.'* (Richardson, 2003a, p. 1)

The points that were raised highlight the importance of how descriptions are worded. For example, the use of affective phrases such as 'the candidate shows enthusiasm for' were criticised for being too subjective to assess. Richardson (2003a) also found evidence of confusion regarding the role of the descriptions, with participants in both the original exercise and the validation study referring to them as 'criteria'. This study highlights how subjective grade descriptions can be. There was clearly ambiguity in the language and a division in opinion between the teacher-examiners who developed the descriptions and the test developers and administrators who were involved in the validation study.

A similar method was also used for a GCSE Applied Business Studies specification (Richardson, 2004). Participants struggled to develop a single list of descriptions that worked for both elements of the assessment; and examination paper and a portfolio. The two forms of assessment provided different contexts making it difficult to make combined and definitive judgements about the overall standard of performance. Richardson (2004) concludes by saying 'it is important to emphasise the judgemental nature of performance descriptors… (they) cannot paint a definitive picture of student performance' (p.6). These studies developed descriptions using a methodology that does not employ statistical information but, despite this limitation, these issues are likely to be common when writing grade descriptions. There is a danger that, whatever methodology is employed, descriptions may confuse or mislead teachers simply because they are trying to communicate something which is abstract and transient.

Another potential issue is that grade descriptions could be actively damaging to learning. This was certainly a concern when grade related criteria were under development. The fear was that providing precise information about what was to be assessed may have an adverse impact on learning by causing teachers to focus too heavily on the criteria and thus to sacrifice educational breadth (Cresswell, 1987b). Specifically, it was noted that it would have appeared reasonable to schools to divide the teaching of a subject's content along the lines established in the criteria. It was feared that this could lead to candidates' having a disjointed experience of the subject as a whole because interlinked areas of content would become isolated from one another (Gipps, 1990). In other words, specifying details of assessment can have a disproportionately strong impact on teaching.

This type of concern remains contemporary. There is evidence that the accountability systems which accompany high stakes testing can have an impact on teacher and student behaviour (see Acquah, 2013 for a summary with an assessment focus). The fundamental issue may be that the interface between curriculum and assessment is unclear in England. The purposes of the two can often become conflated, meaning that assessment drives learning to a greater degree than the curriculum does (see Isaacs, 2014; Stobart, 2008 for more on this issue). It is therefore important that the role which grade descriptions play is carefully considered.

Conclusion: Grade descriptions, in their current guise, are of limited use to teachers

Previous sections have outlined the limitations to how precise and detailed grade descriptions can be. This undermines their usefulness to teachers, who are looking to plot a pathway to success for their students. Further to this, it is reasonable to state that qualitatively defining the performance expected at a given grade is inherently challenging. Grade descriptions must be interpreted by the reader and therefore their meaning can be misconstrued. Such an issue is particularly problematic when it can have a direct impact on teaching and learning. Given these concerns, there is an argument to be made for abolishing grade descriptions. Stakeholders are keen to grasp the performance standards required for each grade but given that these very standards are transient, this may not be possible. A more fruitful approach may be to better communicate the manner in which standards are set and maintained for GCSEs.

Empirical approaches to developing grade descriptions – Exemplifying performance in context

Once the new GCSEs are firmly established and empirical evidence about actual candidate performance is available, the role which grade descriptions play can change. The emphasis could shift to post-hoc approaches that describe which elements of the assessment emerged to differentiate between grades. Deriving grade descriptions from empirical data can be achieved in one of two ways. Firstly, the process can be a largely qualitative one that draws on the expertise of assessors and educationalists to articulate the performance they have seen at each grade (e.g. Richardson, 2004). Alternatively, the approach can be more quantitative. Item level data from a given assessment can be used to identify questions which discriminate effectively between candidates on either side of a component grade boundary. Once these items have been selected the nuance of the how the performances at each grade differs can be described qualitatively by subject experts (e.g. Greatorex et al., 2001). **Appendix B** provides a more detailed description of one such approach. Empirical approaches have the advantage of being rooted in genuine candidate performance, allowing the resultant descriptions to serve as tangible examples of differing performance profiles. They are contextualised because they represent the actual performance of candidates on actual examination items.

The downside to such post-hoc empirical approaches is that they are difficult to generalise. The approach is based on a specific examination and therefore the resulting grade descriptions refer only to performance by a given cohort on a given examination paper. In addition, the particular paper from which the analysis and descriptions are drawn may not necessarily have elicited the type of discrimination that the test developers were looking for. The paper may fail to adequately represent the overarching objectives for the qualification. In short, grade descriptions developed using this methodology are indicators of *past* performance rather than aspirational indicators of what the assessment is designed to measure. Greatorex (2005) captures this neatly:

> *Writing grade descriptors grounded in empirical evidence is arguably an improvement upon methods which are based upon examiners' expectations alone. However, one problem of using experimental evidence in this way is that it is a post-hoc method with grade descriptors based on what candidates achieved rather than on what the examinations were designed to assess* (Greatorex, 2005, p. 10).

Despite this, there is evidence that grade descriptions developed using this methodology in one examination series are usually valid in subsequent series (Greatorex et al., 2001; Greatorex, 2002). Greatorex (2001) found that the descriptions were not applicable to *every* single script or item but in general they were still considered valid by examiners. One issue to consider is whether or not this methodology would be appropriate for all assessments and subjects. There is some evidence to suggest that the approach is effective for numerical, written, short answer and essay questions (Greatorex, 2001), which implies that the methodology could be broadly applied. However, additional research which explores the validity and appropriateness of the approach for a variety of different item types in a variety of different subjects would be necessary if it were to be confidently applied across a suite of new specifications.

Arguably, grade descriptions do not have to be comprehensive to be useful. As Greatorex (2001) states: 'grade descriptors are indicators of performance, not lists of requirements; in a disclaimer model, not all characteristics will be seen in the scripts' (p.463). This might be perceived by teachers as vague but it is important that such limitations are made clear. Grade descriptions are not the basis of grading and should not be presented in a way where they could be misconstrued as such. If their importance with regard to marking and awarding is over-emphasised then descriptions are likely to have an inappropriate impact on how teachers prepare their students for assessment, which, as has been discussed, may have undesirable consequences for teaching and learning. Empirical post-hoc methods reference specific examination papers and so the descriptions that arise constitute *exemplification* of the performance standard at a given point in time. This type of description is arguably more appropriate as a reflection of how grades are actually awarded.

Discussion

With the reform of GCSEs comes an opportunity to refine the way in which awarding bodies and the regulator describe and discuss performance standards. Ofqual have recognised the concerns of stakeholders regarding the first few years of the new specifications and are planning to provide grade descriptions that broadly communicate the performance standards that are likely to relate to the new grades (Ofqual, 2014). However, this paper has argued that the nature of the assessment system in England imposes fundamental limitations with regard to how precise grade descriptions can be. Arguably, these limitations are such that grade descriptions cannot serve a purpose during awarding and are unlikely to be useful to teachers.

It is worth considering whether or not grade descriptions should be provided in the future. As the new specifications become established, grade descriptions are likely to decline in relevance because stakeholders will have concrete examples of what performance at each grade looks like. The consultation suggests that stakeholders are generally keen to garner a range of feedback in order to better understand performance standards:

> *The most mentioned request for further information in addition to grades was that information is provided on student performance on a question by question basis or grades for individual components within the overall grade for the subject.* (YouGov, 2013, p. 80)

At the moment, for a fee, teachers and candidates can request that their examination scripts are returned and they can gain additional feedback through enquiries after results. As the use of technology, particularly with regard to e-marking, becomes increasingly prevalent in assessment, teachers can also be provided with an array of statistics about their students' performance and the performance of the cohort in general. This type of data could be used to provide exemplification of the performance that is required for candidates who are seeking to achieve specific grades.

Empirical post-hoc methods, such as that outlined in **Appendix B** (Greatorex et al., 2001), could be used to provide such exemplification. They can offer a systematic approach to capturing the type of performance that has discriminated between candidates at different grades in the past and they are rooted within the context of examinations. The output would have an appealing transparency and could be 'rebranded' to reflect the fact that, methodologically at least, it is better defined as grade-related performance exemplification than grade description. Such exemplification could be combined other information to provide teachers with what would, arguably, be a much more satisfactory understanding of performance standards; one that acknowledges the nuances of assessment and awarding while providing concrete and contextualised examples at a variety of standards.

References

Acquah, D. (2013). *School Accountability in England: Past, Present and Future.* Manchester: Assessment and Qualifications Alliance.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., … Wittrock, M. C. (2014). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's*. United Kingdom: Harlow: Pearson.

AQA. (2012). GCSE Maths Specification. Retrieved from http://filestore.aqa.org.uk/subjects/AQA-4360-W-SP-14.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.

Baird, J.-A., & Scharaschkin, A. (2001). *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*. Manchester: Assessment and Qualifications Alliance.

Cresswell, M. J. (1986). *Grade Criteria - Some General Issues*. Manchester: Assessment and Qualifications Alliance.

Cresswell, M. J. (1987a). Describing Examination Performance: grade criteria in public examinations. *Educational Studies*, *13*(3), 247–265.

Cresswell, M. J. (1987b). *Grade Criteria: Developing Performance Matrices Guidance Paper For Syllabus Working Groups*. Manchester: Assessment and Qualifications Alliance.

Gipps, C. (1990). *Assessment: A teachers' guide to the issues*. London: Hodder and Stroughton.

Gove, M. (2013a, January). Ofqual policy steer letter: Reform of GCE A Levels. Retrieved from http://media.education.gov.uk/assets/files/pdf/l/ofqual%20letter%20alevels%20v2.pdf

Gove, M. (2013b, February). Ofqual policy steer letter: reforming Key Stage 4 qualifications. Retrieved from http://www.education.gov.uk/schools/teachingandlearning/qualifications/gcses/a00221366/gcse-reform

Greatorex, J. (2001). Making the Grade - How Question Choice and Type Affect the Development of Grade Descriptors. *Educational Studies*, *27*(4), 451–464.

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.

Greatorex, J. (2003). Developing and Applying Level Descriptors. *Westminster Studies in Education*, *26*(2), 125–133.

Greatorex, J. (2005). A review of research about writing and using grade descriptors in GCSEs and A levels. *Research Matters (Cambridge Assessment)*, *1*, 8–11.

Greatorex, J., Johnson, C., & Frame, K. (2001). Making the Grade--Developing Grade Descriptors for Accounting using a Discriminator Model of Performance. *Westminster Studies in Education*, *24*(2), 167–181.

Isaacs, T. (2013). The diploma qualification in England: an avoidable failure? *Journal of Vocational Education & Training*, *65*(2), 277–290.

Isaacs, T. (2014). Curriculum and assessment reform gone wrong: the perfect storm of GCSE English. *The Curriculum Journal*, *25*(1), 130–147.

Jones, B. (2009). *Awarding GCSE and GCE - Time to Reform the Code of Practice?* Manchester: Assessment and Qualifications Alliance.

Massey, A. J. (1982). Assessing 16+ Chemistry. The exposure - mastery gap. *Education in Chemistry*, (September), 143 – 145.

Nash, R. (2005). A change of direction for NCEA: on re-marking, scaling and norm-referencing. *New Zealand Journal of Teachers' Work*, *2*(2), 100–107.

OCR. (2013). GCSE Maths Specification A. Retrieved from http://www.ocr.org.uk/Images/83249-specification.pdf

Ofqual. (2013). Setting standards. Retrieved from http://ofqual.gov.uk/standards/summer-exams-2013/setting-standards/

Ofqual. (2014). Ofqual consultation on Setting the Grade Standards of new GCSEs in England. Retrieved from http://comment.ofqual.gov.uk/setting-the-grade-standards-of-new-gcses-april-2014/

Perie, M. (2008). A Guide to Understanding and Developing Performance-Level Descriptors. *Educational Measurement: Issues and Practice*, *27*(4), 15–29.

Richardson, M. (2003a). *Validation of Grade descriptors for World Class Tests*. Manchester: Assessment and Qualifications Alliance.

Richardson, M. (2003b). *Developing grade descriptors for World Class Tests for nine-year-olds*. Manchester: Assessment and Qualifications Alliance.

Richardson, M. (2004). *Performance descriptors for GCSE Applied Business Studies - A report on the development of performance descriptors for Summer 2004 awarding meetings*. Manchester: Assessment and Qualifications Alliance.

Sadler, D. R. (1987). Specifying and Promulgating Achievement Standards. *Oxford Review of Education*, *13*(2), 191–209.

Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. New York: Routledge.

Whetton, C. (2009). A brief history of a testing time: national curriculum assessment in England 1989–2008. *Educational Research*, *51*(2), 137–159.

Wyse, A. E. (2013). Construct Maps as a Foundation for Standard Setting. *Measurement: Interdisciplinary Research and Perspectives*, *11*(4), 139–170.

YouGov. (2013). GCSE reform: analysis of consultation responses. Ofqual. Retrieved from http://ofqual.gov.uk/ofdoc_categories/consultation-docs/

Appendix A: Grade descriptions for GCSE Mathematics (AQA, 2012)

Grade	Descriptor
A	Candidates use a wide range of mathematical techniques, terminology, diagrams and symbols consistently, appropriately and accurately. Candidates are able to use different representations effectively and they recognise equivalent representations: for example, numerical, graphical and algebraic representations. Their numerical skills are sound, they use a calculator effectively and they demonstrate algebraic fluency. They use trigonometry and geometrical properties to solve problems.
C	Candidates use a range of mathematical techniques, terminology, diagrams and symbols consistently, appropriately and accurately. Candidates are able to use different representations effectively and they recognise some equivalent representations: for example, numerical, graphical and algebraic representations of linear functions; percentages, fractions and decimals. Their numerical skills are sound and they use a calculator accurately. They apply ideas of proportionality to numerical problems and use geometric properties of angles, lines and shapes.
F	Candidates use some mathematical techniques, terminology, diagrams and symbols from the Foundation tier consistently, appropriately and accurately. Candidates use some different representations effectively and can select information from them. They complete straightforward calculations competently with and without a calculator. They use simple fractions and percentages, simple formulae and some geometric properties, including symmetry.

Appendix B: The Greatorex method of developing grade descriptions using empirical evidence

Greatorex, Johnson, & Frame (2001) outline a method for developing grade descriptions which are rooted in the performance of candidates under exam conditions. It is based on an approach which was originally developed by Massey (1982) to produce grade descriptions using empirical evidence for candidates' work in CSE/GCE chemistry. This process has been applied successfully in the context of accounting (Greatorex et al., 2001; Greatorex, 2002) and economics (Greatorex, 2001). There are essentially three stages, as specified below:

Step one: Mastery level analysis - Identifying items that differentiate between grades

For the first step of the process, item level data are used to identify questions which discriminate effectively between candidates on either side of each component grade boundary. To achieve this, candidates are divided into groups based on the final grade that they achieved for a given assessment and the performance of these groups is compared for each item. A grade group that has achieved a specified success rate for an item is deemed to have 'mastered' that item (and the skill underpinning it). For example, let us say that 83% of grade A candidates receive marks for 'Item X' while only 31% of grade B candidates receive marks for the same item. Even fewer candidates with a grade C, say 15%, receive marks for Item X. In this scenario, the grade A candidates are deemed to have mastered Item X while the grade B and C candidates are deemed not to have mastered it. Item X therefore seems to encapsulate a difference in the performance of candidates at the boundary between grade A and grade B. Massey (1982) describes this process as a *mastery levels analysis*. The exact performance level that constitutes mastery can be varied depending on the item - for some items a threshold of 80% may be sufficient but for others it may be inappropriate. It is also important to be cautious of items where thresholds are not aligned. For example, if grade B candidates failed to achieve mastery with regard to a given item while grade C candidates did achieve mastery. Such items would probably not be appropriate for discriminating between grades.

Step two: Kelly's Repertory Grid – Defining the performance differences that were identified

The second step of the process is to qualitatively describe the differential performance between grades that has been identified statistically. Using the items identified in the mastery level analysis, examiners scrutinise the content of the question and the responses in order to identify the knowledge or skill which differentiates the candidates who achieved mastery from those who did not. Responses are placed on a continuum of performance (from best to worst) using Thurstone Pairs – a series of direct comparisons between pairs of scripts. These pairs are analysed using an adapted version of Kelly's Repertory Grid (KRG) such that participants describe similarities and differences between the item level responses from candidates on the grades adjacent to each boundary. These descriptions form the basis of the grade descriptions, which are developed directly from the KRG.

Step three: Validation – Confirming the validity of the grade descriptors

A validation process is necessary to ensure that the grade descriptions remain broadly appropriate across assessments and cohorts. This process has usually been an informal check of the face validity of the grade descriptions, facilitated though their use during the awarding process (Greatorex et al., 2001). Given that the grade descriptions are developed directly from a specific assessment and based on the performance of a specific cohort, this validation is important for ensuring generalisability.