Report for Qualifications Wales

Features of Effective Mark Schemes in Knowledge-based Qualifications: Evidence from a Literature Review and Expert Interviews

April 2018

Angus Alton, Karen Brown and Sarah Maughan

Contents

1 Glossary

Analytical mark scheme: a form of levels-based mark scheme where the knowledge and skills targeted by the item are judged discretely. This usually involves creating a separate set of mark bands for each Assessment Objective (AO), with a description for each and the marker deciding on the best fit of a response for each strand. Such a scheme is often presented in the form of a grid. This grid allows for the marks awarded explicitly to reflect the intended weightings across the AOs.

Assessment objective (AO): a statement of an aspect of a subject that will be assessed in a qualification. It is given a weighting (often within a range) in the subject content defined within regulation. AOs usually cover knowledge and understanding, application, and analysis and evaluation, as well as subject-specific aspects. Each specification sets out how the overall assessment meets these requirements. This helps address the issue of comparability between qualifications from different Awarding Bodies within and across years.

Comparative judgement: an approach to ranking work that doesn't use a scoring system. The markers, who are subject experts, usually explore and agree the qualities they are looking for in a student's work which becomes, in effect, the mark scheme (although this is sometimes assumed to be sufficiently shared or captured by the AOs) and then judge pairs of students' work, simply deciding which is better. Provided there is a sufficiency of such judgements, it is possible to create a student rank order that is at least as reliable as that achieved by normal scoring methods.

Error Carried Forward or Follow Through marks: An approach designed to meet the principle that a student should be penalised only once for any given error. The basic idea is that an incorrect answer later in a question is marked correct if it is correctly calculated using an incorrect answer from earlier in the question. The mark scheme will have to make clear any limitations in applying the system, for example, the nature of the error.

Exemplars: Real examples of student scripts, marked and often with a commentary, used to illustrate the mark scheme. Not to be confused with a model answer.

Halo effect: this describes an effect whereby a marker finds it hard to keep a pre-existing judgement (favourable or unfavourable) about a student separate from the specific judgement required. It is an effect that is considerably mitigated by on-line marking done on limited sections of a paper. It is particularly likely to play a role when using an analytical mark scheme where markers have to make several independent judgements of a piece of work. It has also been found to have an effect when assessing spelling, punctuation and grammar or quality of language.

Holistic mark scheme: a form of levels-based mark scheme in which the marker is required to judge the answer against the full range of what the item is targeting in a single step. The performance descriptions therefore have to indicate what is expected in each of the targeted AOs and the marker has to make a decision of best fit across all the AOs, taking into account any differences in weighting.

Indicative content: examples of the type of material a student may include when answering a question. It may suggest ways in which the material may be handled differently by students of different attainment. It is usually accompanied by a warning that it is for illustrative purposes and does not reflect required content.

Levels-based mark scheme also called a Levels of response mark scheme: a mark scheme where the marker compares the student's performance with a set of descriptions of performance across the knowledge and skills targeted by the item. The marker has to decide which description best fits the student performance and then awards a mark from within a band of marks it applies to. Such schemes can be used on relatively low-tariff items requiring qualitative assessment to much more extensive tasks. They are characterised as either Holistic or Analytical.

Model answer: the examiner writes a model of how the student is expected to respond.

Outcome space: the complete set of possible responses to a given task. This includes both the full range of rewardable material and that which does not receive credit. The mark scheme is essentially the anticipated outcome space, and the aim is that this should perfectly match the actual one. The more tightly defined the task, the easier it is to control the outcome space.

Points-based mark scheme: a mark scheme which lists the points that are acceptable for an answer. This will normally award one mark per acceptable point up to a given maximum, and there are usually a few more acceptable answers than the maximum mark. Such schemes can be used for items carrying just a couple of marks to assessing an essay. Such a scheme will also need to indicate any variations in the expression of each acceptable point which will be allowed.

Seeds: examples of live students' work which have a mark agreed in advance by senior markers and which are used to monitor marking accuracy during the marking process. This approach is only applied for on-line marking. Where it is paper-based, **standardisation scripts** are used

Standardisation meeting: a meeting of all the examiners led by the Principal Examiner designed to ensure that all markers are marking consistently. Markers will normally mark a number of scripts or items, covering as much of the mark range as possible, where the marks have been agreed in advance by senior markers. The meeting will review and discuss the marks awarded. The meeting is also an opportunity to review the mark scheme in the light of real responses and clarify or amend its requirements as necessary.

Standardisation scripts: where marking is done on paper, immediately after the **standardisation meeting** all examiners mark a set of scripts and have their marking reviewed by a senior marker and it has to be approved as sufficiently accurate before they are allowed to commence live marking. Nowadays, these scripts are usually the same set of photocopies for everyone, with previously agreed marks, and the criteria for what is sufficiently accurate are also defined.

2 Executive Summary

In support of the evaluation, approval and continuous improvement of qualifications, Qualifications Wales commissioned AlphaPlus Consultancy to research and produce guidance on the features of effective mark schemes for knowledge-based qualifications. This report summarises the findings from: a literature review, stakeholder views from assessment professionals from Awarding Bodies and UK Qualification Regulatory Authorities, specialists in assessment within specific subjects; and an expert practitioner focus group.

The principles of designing good mark schemes stem from their purpose, which is to support valid and consistent assessment, across a range of evidence, to inform conclusions on grading. To achieve this a mark scheme must define the scope of acceptable responses and the individual criteria for awarding marks against responses and how this relates to grading decisions. A mark scheme must inform markers sufficiently so that the conclusions are reliable and consistent across markers.

There was consensus from the literature review, stakeholders and the expert practitioners that to perform as intended, a mark scheme must be closely linked to the overall qualification assessment specification and intended outcomes. Furthermore, it should be developed in conjunction with the assessment questions, to ensure consistent alignment throughout all elements of the qualification. It was also accepted that the areas where mark schemes face significant challenges are often associated with higher tariff, open questions which can generate a wide range of responses. There was clear consensus that there are valid differences in approach for some subjects and therefore there is a clear need to retain variations in mark scheme design, to meet the required outcomes for different qualifications.

Key features included in good mark schemes were concluded to be:

- Sufficient detail on anticipated and required answers and indicative content to inform consistent decisions, across all available marks
- Clear guidance on how to deal with unexpected answers, focussing on the principles behind required outcomes
- Avoidance of the use of single model answers and instead use of real student responses as exemplars in the standardisation process
- Clarity of command words leading from AOs and links from these to the mark scheme and question design
- Clarity of and clear distinction between both level and mark band descriptors
- Avoidance of additional requirements (hurdles) beyond those specified by the question
- Avoidance of information overload for markers
- Clear and consistent presentation and language across the whole qualification
- Supported by and intrinsically linked to robust and open standardisation and quality assurance processes.

Other issues of note included observations of the increased use of mark schemes to inform the delivery of course content and the inherent difficulties this can cause in terms of restricted focus and teaching to the task, rather than teaching to cover the qualification aims and objectives. Also, the role of comparative judgement in marking is as yet unusual, but if combined with the key features highlighted above, could offer a positive alternative to current models.

Finally, the advent of on-line marking has some implications for mark schemes both in terms of structure and presentation. It is important that Awarding Bodies consider how best to exploit the logistical benefits that arise from marking on-line while recognising the additional complications that may arise.

There was clear support from all strands of the research for further practical guidance to be developed on the design and development of mark schemes.

3 Introduction

Qualifications Wales is the independent organisation responsible for regulating non-degree qualifications, and the qualifications system, in Wales. As part of its qualifications reform work, Qualifications Wales is evaluating and approving new qualifications for use in schools and other educational settings in Wales. Based on this work to date, Qualifications Wales has concluded that the on-going qualifications reform process would be supported by research and stakeholder engagement on the general principles of effective mark scheme design in knowledge-based qualifications and the production of a good practice guide informed by the findings.

AlphaPlus Consultancy have been commissioned to undertake this project for Qualifications Wales.

4 Methodology

This report presents the findings from three strands of research which will be used to inform the development of a good practice guide on mark scheme design. The three strands were:

- A literature review
- Stakeholder interviews
- A focus group with AlphaPlus staff.

4.1 Literature Review

The literature review focused on the principles of effective mark scheme design. It focused particularly on the research questions:

1. *What does the academic literature tell us about the principles of good mark scheme design?*
2. *What features of the mark scheme support accurate marking and valid assessment?*

The search included literature from 2000 onwards, published in any jurisdiction and published in English. Where possible, research literature was selected which focused on the assessment of 14- to 19-year-olds either in general or vocational education.

The literature review followed the stages set out below:

- Identify suitable sources/databases, and establish/refine relevant search terms
- Undertake the literature search and produce a 'long-list' of relevant references
- Review the long-list and refine this down into a short-list of high quality articles
- Summarise the literature using an agreed template
- Use the summaries to write the literature review.

In the end a 'long-list' of 24 articles was created. These were evaluated for direct relevance, for whether they provided information about mark scheme design, the features of effective mark schemes, knowledge-based subjects, and if they were based on high quality evidence. From that initial long-list of sources, the most relevant literature was selected to form a short-list of 12 articles to be fully reviewed and summarised.

4.2 Stakeholder Interviews

A list of potential stakeholders to include in the interviews was provided by Qualifications Wales. This was discussed and refined to produce a final list including:

- Senior leaders, subject experts and researchers from the main awarding bodies in Wales and England
- Subject experts with an involvement in qualifications reform in Wales
- Representatives from the regulators in England and Northern Ireland.

Those on the list were contacted by the AlphaPlus team to arrange a convenient time. The intention was to hold individual interviews where possible, but flexibility to hold a group interview was given for organisations providing multiple staff members.

A semi-structured interview schedule was drafted and agreed with Qualifications Wales (See Appendix 1). This included prompts around the following issues:

- Principles of what makes a good mark scheme
- Different styles of mark scheme
- Subject-specific issues in mark scheme design
- What they would like to see in a mark scheme guide
- The range of constructed response item types to be covered
- The expected level of (mark/re-mark) reliability.

The interviews were conducted using the schedule, slightly varied according to the particular interests/expertise of the interviewee. Interviews lasted about 40 minutes. In the end, we conducted one group interview (with four participants) and 13 individual interviews.

4.3 AlphaPlus Focus Group

In addition to the formal data collection described above, Qualifications Wales was keen to include insights from the senior team at AlphaPlus. Various members of the team have conducted mark scheme research and development in the past which had been reflected in the proposal for the work. Qualifications Wales requested a way of including this expertise in the data collection. A focus group was conducted with the available team members. This was led by an expert associate and included three AlphaPlus Directors.

The report that follows summarises each strand of evidence collection and draws together the key findings from the different strands in the Discussion section.

5 Summary of Literature Review

The common premise across the literature reviewed is that: measuring the extent of someone's knowledge or ability should be determined by their attainment and not be constrained by deficiencies of the exam or associated mark scheme. In this context, the concept of producing a "good" mark scheme hinges on the ease with which it can be applied consistently and clearly across a range of answers, allowing markers to distinguish between responses in agreed increments.

*"An effective rubric has a well-defined list of criteria for the test-takers to know what is expected of them and for the raters to be able to properly evaluate the responses... contains standards of excellence for the different levels of performance…(*and*)… has gradations of quality, or a scale, based on the degree to which the standard has been met." (*Ghalib & Hattami, 2015 p.226[1]*)*

At the heart of this concept is the anticipation of answers, whether this is done through:

- Extensive mapping of answers across a wide "outcome space" representing the range of responses that might be produced and distilling this to focus on the expected answers (Ahmed & Pollitt, 2011, Tisi et al, 2013)
- Through the provision of exemplification of "model answers" and articulation of poor responses (Ahmed & Pollitt, 2011)

[1] NB References are provided in Appendix 2.

- Creating specific criteria for grading the quality of a response on a more holistic level (Ghalib & Hattami, 2015).

Currently, there is no single regulatory prescription for mark scheme design and therefore mark schemes have been developed by a variety of awarding bodies which vary both in their style and approach. These range from highly structured and definitive, short answer schemes, to more open ended holistic schemes often associated with more extended response questions, and often commonly a mixture of both approaches (Ofqual, 2014).

The literature review revealed consensus that it is easier to be accurate and consistent when using a highly specified mark scheme, with lower maximum marks and for questions targeted at lower attainment grades (Tisi et al, 2013, Black et al, 2011, Ghalib & Hattami, 2015). Developers may therefore conclude that it would be preferable to focus purely on this type of question and associated mark scheme. However, to do so would miss the fundamental point that the nature of a mark scheme should be chosen in response to the outcome it hopes to capture, including covering an appropriate complexity of questions. Some evidence was presented to suggest that a purely analytical approach to marking is more reliable than a holistic approach (Tisi et al, 2013, Ghalib & Hattami, 2015), however, with reference to this fundamental point, no clear evidence was presented to suggest either approach is more valid across all types of outcome (AlphaPlus, 2014, Bramley, 2008, Ghalib & Hattami, 2015). In other words, it is correct to acknowledge that different needs will require different approaches in mark scheme design.

The judgement on the validity of a mark scheme can be captured in three questions. Namely, whether the mark scheme:

- supports assessment of the intended outcomes of the question(s)
- provides support to markers, particularly when faced with unexpected answers
- provides the means for a consistent response across markers.

Taking these three points in turn, some conclusions and suggestions arise from the literature review which could be used to support those developing mark schemes.

5.1 Supporting the assessment of intended outcomes

Repeated references are made, across much of the literature, that it is neither helpful nor possible to separate the development of a mark scheme entirely from the development of the question papers and from the purpose and content of the qualification itself (AlphaPlus, 2014, Pinot de Moira, 2013, Christie et al, 2015). In fact, the argument goes further: that the approach used for a mark scheme should only be decided after establishing that the specified objectives of the qualification are clear and whether the challenges posed through questions truly reflect the outcomes required. Only then can mark schemes be devised which relate to those same outcomes (Ahmed & Pollitt, 2011, Christie et al, 2015).

A simple example of this could be the development of a mark scheme for a multiple-choice assessment, which may at first glance appear as a straightforward task, but in fact is entirely reliant on the focus of the questions on the identified objectives and the usefulness of the distracters in maintaining that focus (Ahmed & Pollitt, 2011). Similarly, the mark scheme must ensure that a student's response is evaluated in terms of the intended focus of the question rather than rewarding more general knowledge and understanding or examination technique (Tisi et al, 2013).

Issues become more prevalent when the development of the mark scheme appears to misalign with the intended outcome of the questions. Examples of this issue (Tisi et al, 2013, Ahmed & Pollitt, 2011, AlphaPlus, 2014, Ofqual, 2014) are:

1. When a question asks a student to provide an explanation of a concept, or set of data, but the mark scheme is based purely on a points-based recognition of content marks rather than the demonstrable explanation
2. When there is a mismatch in the level of skills and understanding required by question marks and mark schemes, for example, if additional bonus points are offered for issues not referenced in the question.

There is consensus from the literature that: the number and distribution of levels and marks within a mark scheme has an impact on the consistency of mark allocation. This consensus falls into two common areas:

1. The ability of markers to clearly understand the weighting of responses in relation to individual mark bands/levels (Pinot de Moira, 2011)
2. The ability of markers to offer the full range of marks across a question, particularly in the highest extremes of higher tariff questions.

In both cases, developers are referred first to the importance (weighting) of the item in terms of the qualification outcomes, and then to the extent to which the levels and the individual marks can be uniquely described and distinguished (Black et al, 2011, Bramley, 2008, Ofqual, 2014, Pinot de Moira, 2013).

A secondary issue regarding number and distribution of levels which is well documented is that a low tariff item is far more likely to be marked accurately than one with a higher tariff as the number of decisions the marker has to make is smaller (Bramley, 2008, Black et al, 2011). Items with a higher mark range are likely to have fewer responses categorised in either of the extremes of the tariff which can mean that the marking of these responses is less reliable.

5.2 Supporting markers in reaching conclusions

There is a balance to be struck between designing a mark scheme to cover predicted outcomes and not designing one which restricts answers or dictates outcomes (Tisi et al, 2013, Ahmed & Pollitt, 2011, Black et al, 2011).

> *"…the better the overlap between the observed and expected outcome space…the more reliable the marking will be."* (Tisi et al, 2013 p.24)

The crux of the problem is to design mark schemes which contain enough detail to clearly indicate what the expected outcome is and how marks should be awarded consistently to discriminate fairly between students, whilst remaining generic enough to cover a range of responses (Christie et al, 2015, Ofqual, 2014, Pinot de Moira, 2013).

> *"Marking reliability was higher, …when the band descriptions were generic rather than including indicative content specific to the particular item."* (Pinot de Moira, 2013 p.8*)*

There should be a clear distinction of what is required in order to categorise a response in one band description or another. Where a clear distinction is not made it is possible that this might indicate that there are too many bands to allow for distinction between them, or that the terms used within the bands are not clear enough to guide markers consistently. There was some disagreement whether a specific taxonomy was best to control terms used in mark bands or whether this might introduce unnecessary regulation. A useful "rule of thumb" is that, if it is not possible to give a clear and separate definition of the terms used in level descriptors/mark bands, this might indicate that the bands overlap or that the terms should be revisited (Bramley, 2008, Ofqual, 2014).

> *"Are the standards' descriptors explicit, devoid of subjective words, and positively worded in terms of what students must do?"* (Christie et al, 2015, p.30)

One contentious issue covered by the literature is the provision of detail on poor responses as well as good/model responses. The conclusion is that, especially where multiple outcomes are expected, this is useful to provide better understanding of how to distinguish between responses, but that this can overload markers with information and increase the cognitive demand placed on them; it is likely to be more effective for the mark scheme to define the overarching principle than to list multiple possible responses. Care should also be taken that in overly defining answers, this does not encourage negative marking. (Ahmed & Pollitt, 2011, Bramley, 2008, AlphaPlus, 2014)

> *"For the points-based items, the presence of qualifications, restrictions and variants seemed to increase the level of agreement very slightly…For the objective items [items with a single correct answer], on the other hand, the presence of qualifications, restrictions and variants seemed to reduce the level of agreement."* (Bramley, 2008, p.12)

A related point concerns the spread of marks across bands. Statistical evidence was largely inconclusive but erred slightly towards the suggestion that marking reliability is increased if marks are spread evenly across the bands (Pinot de Moira, 2011, Pinot de Moira, 2013). This is possibly down to the increased simplicity of this approach for markers to deal with.

On this note, several of the studies concluded that one of the keys to developing a "good" mark scheme was in keeping both mark scheme itself and the guidance to markers clear, simple, uncluttered and consistent. There appears to be a direct correlation between increased cognitive demand on markers e.g. through multiple options to accrue marks or inconsistency in approach across a whole mark scheme and an increased inconsistency of marking (Child et al, 2015, AlphaPlus, 2014, Bramley, 2008, Pinot de Moira, 2013).

> *"…visual comparison of the generic and specific levels-based mark schemes reveals … (that) … Generic mark schemes are often simple, neat and uncluttered. They are the same in format throughout the unit and, therefore, require less cognitive demand of the user. Levels-based mark schemes which include indicative content in the bands tend to be lengthier and, by definition, differ across the unit."* (Pinot de Moira, 2013 p.8)

5.3 Supporting consistency of response across markers

It is important to acknowledge that levels-based marking, whether analytical or holistic, poses the most difficult challenge in terms of providing marker guidance and encouraging consistency. This appears to be due mainly to the fact that it is often associated with those questions accepted as being the hardest to mark (high tariff, open/unconstrained, maximum space and time for response, i.e. the longer the answer the more there is for the marker to evaluate in terms of the mark scheme) (Ahmed & Pollitt, 2011, Black et al, 2011, Bramley, 2008).

> *"…it is a higher quality of response rather than a greater item difficulty which lowers reliability"* (Pinot de Moira, 2013)

> *"…some high-performing students could receive lower marks because they are more likely to give unexpected answers the mark scheme does not capture"* (Ofqual, 2014, p.38)

As highlighted in the opening section of this review, the decision to reduce the use of this type of question can have significant impact on the opportunity for students to demonstrate more creative thought processes without jeopardising the demonstration of content-related knowledge (Lille & Romero, 2015).

> *"… the use of questions with longer responses is an important part of the assessment process and so an educational system may choose to accept the lower levels of reliability."* (Tisi et al, 2013 p.2)

Even with mark schemes and guidance in place, a more open, holistic approach to marking can be more influenced by marker "expert opinion" than in more restricted mark schemes. For example, markers may reach

the same conclusion on final mark, but when asked to justify their conclusions may have very different reasoning. Where this is underpinned by a focus on the quality and relevance of the response and valid differences of expert opinion, this is deemed acceptable (Black et al, 2011), but where this is the result of inexperience, a lack of/poor guidance in the mark scheme, for example relating to the use of students' own formulae/phrases, the inconsistency is not acceptable (Black et al, 2011, Tisi et al, 2013). The point taken from this conclusion is that mark schemes should be developed to support the lowest denominator of marker e.g. the new/inexperienced marker.

Consideration should be given to understanding and reflecting the processes undertaken by markers when designing a mark scheme. This topic is wide enough that it could become the subject of a review in its own right, but could be summarised as markers using five strategies: matching, scanning, evaluating, scrutinising and no response, to check what they found in an answer against the mental model they have formed of the mark scheme, often looping through the process several times before reaching a mark conclusion. (Suto and Greatorex (2008), AlphaPlus, 2014).

The impact of this is that the more open the response, the higher the level of complexity of approach required to mark an item (Bramley, 2008).

The research covered by this review also suggests that simple changes to the presentation of mark schemes and associated guidance can increase usability and therefore reliability and consistency (Child et al, 2015, Bramley, 2008, Ofqual, 2014, Pinot de Moira, 2013). These included:

1. The order of presentation of levels: lowest first or highest first, can have a bearing on a marker marking positively or negatively. There is some evidence[2] to suggest that there is better reliability for mark schemes where the marks are described in ascending order (Pinot de Moira, 2013)
2. The presentation of levels in a grid-like format to separate the evaluation of AOs can be useful to clarify different bands, but carries the risk of over specification and complexity as the number of categories increases (Pinot de Moira, 2013)
3. There is some evidence that small changes to the formatting of the mark scheme can mitigate against a lack of experience for example, by making changes to the proximity of guidance related to specific questions, the bolding of key terms, and the page formatting (Child et al, 2015, Christie et al, 2015)
4. Other points were raised concerning the design of question papers e.g. the influence of space allocated for an answer and therefore the amount of script an examiner might need to review; the response to diagrams and tags, etc. was found to have a higher agreement for markers than full sentence responses. These points are considered to be outside the scope of this report, but should be considered in the development process (Bramley, 2008).

One specific point noted on the content of mark schemes concerns the assessment of the quality of written communication. This was not addressed in many of the studies but, where it was mentioned, the conclusion was that for the majority of content-driven subjects this should be assessed as a separate outcome rather than within question responses (Pinot de Moira, 2013, Ahmed & Pollitt, 2011). This is because the AOs within a subject rarely include quality of communication, even implicitly. As a result, the accuracy of marking focuses on a particular skill or area of knowledge which may be affected if markers have to include quality of communication or spelling, punctuation and grammar (SPaG) within the mark.

[2] Findings passed the five per cent threshold for significance, but only by a small amount.

Several of the reports considered the importance of the standardisation meetings and quality assurance processes which underpin any mark scheme. The consensus is that these approaches are considered vital in terms of honing and improving the mark scheme particularly in the early stages of marking before wider use (AlphaPlus, 2014, Ofqual, 2014).

References were also noted to mark schemes being used to inform the delivery of courses and this may be influencing the actual content being delivered to either restrict or focus on one aspect of the qualification content (Lille & Romero, 2015).

> *"…for students, assessment criteria are integral to their understanding of tasks and success in undertaking them"* (Christie et al, 2015 p.27)

Whilst it is acknowledged that this is not the primary purpose for mark schemes, developers should be aware of this when creating a mark scheme and particularly when considering the inclusion of indicative content/exemplification and in ensuring that there is clarity and a clear distinction between mark bands.

In final conclusion from the literature, there is consensus across the sources that there has been relatively little research focussed purely on mark scheme design and that the consolidation of guidance on producing mark schemes would be considered valuable to those writing and judging them (AlphaPlus, 2014, Ofqual, 2014, Pinot de Moira, 2013).

6 Summary of Interviews

Interviewees came from a wide range of backgrounds, covering different subject specialisms, roles and experience, but there was a reassuring level of consensus in the responses. This summary broadly follows the structure of the interview schedule (see Appendix 2). However, there were several points raised which covered all forms of mark scheme, effectively becoming principles of good practice. How these principles apply in detail to different item/mark scheme types will be described where relevant, but it is useful to draw the principles out.

One point of note is that the interviewees were well-acquainted with much of the literature, or at least with many of the findings and recommendations arising from it, which may to an extent explain this level of consistency.

First, it is vital that item and mark scheme must work in tandem, being developed at the same time and constantly checked that they remain in line. In fact, it may be better to see the item as the bridge between the specification and student behaviour (captured in the mark scheme). Once there is agreement that the mark scheme represents a given aspect of knowledge, understanding or application within the specification, then it is the item that needs adjusting to ensure that it is going to produce the desired performance. In other words, the mark scheme comes first.

An extension of this is that it is essential that there is absolute clarity about what aspect of the specification an item (or rather its mark scheme) is targeting. Not only does this make it easier to track coverage, but it helps in establishing the principle around which decisions to award credit are made. For this is another point about which there was agreement: the task of marking is cognitively demanding at best, and so clarity is more important than comprehensiveness. In Alastair Pollitt's term, it is vital that the wording of the item is sufficiently tight to control the outcome space in order to make it easy to judge whether or not a response addresses the principle behind the item. To aid this and in the interests of building shared understanding, many interviewees recommended stating explicitly the AO (or part AO) each item is addressing.

There was also unanimity that the mark scheme should not be seen in isolation, but is part of the whole process of standardisation, including the standardisation meeting where markers work on exemplar scripts which enables the Principal Examiner to interpret the mark scheme on real responses. A great deal of care is

taken to ensure that the exemplar work covers the full range of responses including, as far as possible, unexpected ones. At this point, markers also have the opportunity to raise any responses they have come across in their preliminary work which they have found hard to assess. It is quite common for the mark scheme to be amended at this stage.

It was also pointed out that the publication of mark schemes is a significant change to the system. For example, it does not include the training elements that markers receive, increasing the possibility that there will be misunderstandings. (It was suggested that it might mitigate this risk and generally improve public understanding of the system if the published mark scheme was accompanied by marked and annotated exemplar materials, although this would increase the burden for awarding bodies). It was agreed, though, that publication should not be allowed to interfere with the primary function of a mark scheme as a main driver to consistent marking.

6.1 Writing mark schemes for short answer items

There are, in fact, several types of short answer items, each requiring different approaches. Behind all of these, however, is the need to make sure the item makes clear what is expected. In particular, it is important to ensure the command words are used consistently (at least within a subject) and in accord with the expectations of natural language. For example, make sure that an item asking for a description expects a description; one asking for an explanation expects an explanation.

6.1.1 Items (usually single mark, although an item may involve several such points) requiring a specific right or wrong answer

Here what is needed is clear specification of the right answer, together with an equally clear statement of any tolerance, for example, whether mis-spelling is allowed or what degree of accuracy is needed for a numerical value or whether units are needed.

6.1.2 Items requiring a fuller answer, such as a definition or explanation of a specific phenomenon in perhaps one or two sentences or examples of a particular phenomenon

Here the scheme must again show the right answer (here, perhaps, the received definition, with a clear indication of which elements in it are worthy of reward) together with a clear statement of the principle of how to deal with (the majority of) answers which will not match that definition perfectly. This will normally involve deciding whether the response gets the concept across – or the parts of it captured in the mark breakdown. A particular aspect of such items is that the number of possible sources for a mark will often exceed the number of available marks: this is reasonable, but it is important that the one does not exceed the other by much. There are two reasons for this: the first is that having very large numbers of acceptable answers is likely to affect the facility and probably the discrimination of the item. The second is that the greater the number of possible answers, the harder the matching process for the marker.

This type of item also introduces a particular feature where the approach seems to go against the principle of positive marking. It is often the case that a student's answer will contain unnecessary material or material that clearly shows that they do not understand the concept, for instance by contradicting something already stated. This leads to two types of guidance needing to be included in the mark scheme. The first covers 'neutral' information which carries no credit but does not affect the accuracy of a response. For example, a response may go into much more detail than necessary about a particular point, meaning the point will be fully rewarded but with no extra credit for the unnecessary material. The second is where the additional material shows real lack of understanding of the topic (perhaps by contradicting something already included) and here the general approach is that no credit should be given even for the accurate material.

By the same token, in some subjects, responses containing lengthy lifts from a passage which happen to contain the information asked for are not rewarded, nor are those which quote exactly when the question asks

for the student to use their own words or vice versa. It is important here to ensure that there are sufficiently precise alternatives to the words in the passage to require them to be paraphrased.

6.1.3 Items calling for fuller responses (up to 10 marks)

Here, three approaches to the mark scheme may be valid depending on the precise nature of the task. Where an item is primarily knowledge-focussed, for example, describing the principal stages of a process, it is possible to use a points-based method. Here the mark scheme must make clear, as always, the principle on which credit is to be awarded, followed by the rewardable points (which may include prohibited combinations, i.e. where both points are admissible, but only count for one point if appearing together) and any key exclusions.

It is easy to see how such an approach can quickly become complex and lead to error, since it involves a sequence of mechanical processes together with many minor judgemental decisions. There is research evidence that the method is more accurate than levels of response for relatively low tariff items but becomes less so as the tariff increases. It should also be noted that the difference in the mark awarded by two markers using a points-based approach are marker errors, arising from a failure to reward a valid point or rewarding a non-valid one; whereas it is generally accepted that different marks awarded using a levels of response scheme may represent legitimate differences of professional opinion.

The usual alternative is to use a levels of response scheme, especially where the item is asking for analysis or evaluation. This will normally be tightly focussed on the targeted skill – ideally a single skill/AO, since this improves the focus (the principle, again) – and may well include indicative content within the band descriptors. This is because the level of, say, evaluation of a particular tightly defined topic is likely to be predictable in terms of the points covered as well as the evaluative skill shown. However, there is a risk that this indicative content might be interpreted as required material and thus lead to marking error.

A third approach is a hybrid one. Here some of the available marks are awarded on a points basis, with the remainder to reward the quality of the response. This approach is sometimes used in science and other subjects with a high knowledge content.

6.1.4 How should a mark scheme deal with unexpected answers?

There were two main points covered in terms of unexpected answers, beyond attempting to cover as many as possible prior to operational marking through, for example, the standardisation meeting. The first was that the main method was through encouraging markers to refer problematic answers up to a more senior marker or team leader. It was noted that this process was a great deal more effective and efficient where marking is on-line, since this permitted feedback and dialogue almost immediately. The second was that the process of question paper development is to try and ensure that items are sufficiently tightly focussed and clearly worded to make sure the students know exactly what they are being asked to do, and thus minimise the unexpected.

6.1.5 What is the role of exemplification?

It was almost universally argued that appropriate exemplification was a vital tool in clarifying the mark scheme. This put a significant premium on the time-limited process of selecting the various categories of script to be used for standardisation. It was important to cover the full mark range (it is particularly important to give full marks for relatively low tariff questions, since there is a serious risk of capping marks overall) and, as far as possible, some of the unexpected ways of responding. There was also complete agreement that these should be real student responses and that model answers should be avoided.

6.1.6 What is the role of indicative content?

As already noted, the indicative content provides a useful idea of what the setter has in mind as legitimate responses to a task. This is important particularly early in the development process where it provides a kind of sense-check on the relationship between the task and the mark scheme. Providing an idea of the sort of content a task is expected to elicit as well as the skills required enables the setter to demonstrate its fitness for purpose. It is also important in terms of ensuring markers, especially inexperienced markers or those lacking in confidence, understand the purpose of the question, improving the understanding of what is relevant.

However, indicative content presents two key dangers. The first is that it is easy to provide too much of such content, thus swamping the skills the question is designed to test. The key is the term 'indicative': it is neither necessary nor desirable to try to provide a list of all the information that will arise to a subject expert about a topic. In fact, if it is hard to avoid providing large volumes of indicative content for a question, it may be a sign that the task itself is too unfocussed. This is likely to be a particular concern in low-tariff questions, since it suggests that the outcome space is insufficiently defined.

The second danger, especially where the setter has aimed to make the indicative content as comprehensive as possible, is that it comes to be seen as required rather than indicative. This is likely to have a depressive effect on the marks awarded, since students simply won't provide answers that cover the full range of content, while any unexpected but worthwhile content may be overlooked.

6.2 Writing mark schemes for items requiring extended answers

All interviewees favoured using levels of response schemes, predominantly analytical ones, for items requiring extended answers. At the same time several issues were raised which are described below.

Perhaps the most common issue was that of the qualitative words used in the level descriptors. It is important that these align as far as possible with natural language (thus the hierarchy across the mark bands is unarguable). It is also that the meaning of these words is as clear as possible. A particular weakness in this area is where the steps are, crudely, 'does a little bit of x', does a little bit more', 'does some', etc. and it is important to take pains to avoid taking this route.

The analytical approach was seen to offer both theoretical and technical advantages. On the one hand, it controls the cognitive demand of the marking process by requiring a series of discrete judgements. It also ensures that the award of credit mirrors the intended weighting across the AOs as shown in the specification. (Interestingly, it was suggested that one specification had chosen a holistic method precisely to avoid a particular AO receiving more than its intended weight leading to poor teaching and learning.) On the other, it allows for much more focussed level descriptors, dealing only with a single skill at a time. It also controls the number of marks awarded for each AO and therefore within each level, making the process more manageable and reducing the potential for disagreement.

What the analytical schemes don't necessarily do is allow for imbalanced answers, particularly if the answer is skewed towards the higher order skills such as analysis and evaluation, with knowledge perhaps implicit rather than itemised. The holistic approach can allow for this. However, if there is a regular sense of dissonance between the mark achieved through an analytical scheme and an overall impression of the quality of the work, it might be better to reconsider the way the various elements are weighted than to go for a (less reliable) holistic scheme. Holistic mark schemes are also less laborious to apply, since the marker is awarding only one composite score.

One feature that sometimes occurs within levels-based mark schemes is the introduction, intentionally or otherwise, of hurdles. For example, a higher band descriptor may require the presentation of the case for and against a proposition, thus relegating any answer which doesn't do this (or, importantly, is judged not to have done by the marker) to a mark from a lower band, no matter how good the one-sided case is. Or it may be that a band requires three examples of something, again lowering the mark of a response that gives two examples

that are outstandingly analysed. (It is important here that this is less of an issue if the task specifies the number of examples required and so on.)

There is a particular example of this in the mark schemes for GCSE Modern Foreign Languages (MFL) where the best performance band for language quality requires the use of at least three tenses (the criteria refer to a range of tenses, with three or more being a reasonable enough interpretation of 'a range'). However, this means that a piece, although fluent and impressively written and using two tenses as a natural product of its content, cannot get into that band. It was further explained that this has a negative effect on teaching and learning and can create very artificial writing exercises designed purely to meet the requirement.

Preliminary findings of a major piece of research by Ofqual on levels-based mark schemes, reported at the AEA Europe conference in Prague in 2017 , found that the presence of hurdles was by far the most significant contributor to marking inaccuracy.[3] It is worth noting that the use of bold or italics in the mark scheme was also a statistically significant contributor to inaccuracy (this is in direct contrast to the conclusions found in the current literature where this approach was suggested as improving the usability of mark schemes). The device may have had the effect of making the stressed elements seem to be required and thus they functioned as hurdles. (The presentation of indicative content in the form of bullet points rather than prose also impaired marking accuracy: it is easy to see how such presentation could become something of a checklist. It is worth also pointing out here that the same research suggested that whether the mark bands were presented in ascending or descending order had a small and non-significant effect on marking reliability in this particular research.)

There are essentially two main ways of presenting a levels of response mark scheme (although as noted above other presentational features should not be taken for granted). In the first, it is tied to a particular item, so that the performance descriptor grid appears on the same page as the indicative content, perhaps with the indicative content below. It was stressed on several occasions that the two should be kept as separate as possible, even in these circumstances. This is largely because of the issue noted earlier, that the content comes to be seen as required rather than indicative, where it is mixed with the level descriptors.

The other approach is suitable for papers or sections of papers comprising solely extended items addressing the same set of skills (as happens often in history or English literature papers, for example). Here, the grid will apply to all the questions (and probably over time) and will usually be placed on its own at the start of the mark scheme. The indicative content is then placed with each individual question, since it will need to be distinct in each case.

6.2.1 How should a mark scheme deal with unexpected answers?

This is more of an issue for extended answers, since it is harder to keep control over so much outcome space. It is a particularly important issue because it is here that the predictability of a paper may be an issue, and it is also where the use of prepared answers becomes relevant. In the mark scheme, all that can be provided is guidance on the need to mark what is there in terms of its relevance to the actual question asked. However, there is a real risk of over-rewarding pre-prepared answers and answers where the student simply sets down everything s/he can think of about a topic. (Here, too, the spirit of positive marking does not necessarily produce the best outcomes.) It is therefore vital to ensure that relevance is a major element of the qualities to be rewarded so that an answer, however unexpected, can be judged as an answer to the question.

[3] AEA Europe: Association for Educational Assessment Europe. The findings are, as yet, unpublished.

6.2.2 What is the role of appropriate exemplification?

The role of exemplification in extended answers is essentially the same as for shorter answers: to help markers understand exactly what it is appropriate to reward. In this case, the prime focus is on amplifying the key terms in the performance descriptors. An obstacle to this is, of course, the length of the responses making it much harder to cover as wide a range of responses as is desirable, both for those selecting appropriate exemplars and within the time available for training. One suggestion, especially where the mark scheme was common across questions and examination series, was the development of a set of materials from a previous examination, carefully selected to illustrate key features and accompanied by commentaries. Such materials would also, of course, be very useful accompaniments to the published mark schemes.

6.2.3 What is the role of indicative content?

The role of indicative content is also broadly the same as for any other item: it tries to capture exactly what the setter is looking for in responses to the task, and it can therefore serve to validate the marker's own reactions to it. It is vital that the indicative content is not seen as exhaustive or prescriptive, and mark schemes all contain wording along those lines, encouraging markers to reward any other relevant material. Indeed, making the indicative content apposite but relatively brief has two advantages. First, it makes it obvious that it is not exhaustive and thus implies that the markers are being trusted to recognise what is relevant, an important aspect of confidence building and establishing a community of practice. Second, there is less of it to complicate the process of marking by adding to the cognitive load, a factor which should not be underestimated. There is no need for, indeed no role for, model answers.

6.3 Subject differences

All subjects have their own characteristics, often reflected in the nature of their question papers and accompanying mark schemes. Two subjects which stand out in this area are mathematics and modern foreign languages (MFL).

In mathematics (and in the numerical aspects of other subjects) one issue is follow through or error carried forward. The principle is that a student should not be penalised repeatedly for the same error. Thus, an incorrect answer which is used correctly in subsequent parts of the question to calculate an incorrect answer using the right method will receive credit. This is sometimes limited to an extent by the exact nature of the error, but the principle is not in question. This does have the effect of complicating the marking process somewhat, with markers sometimes having to recalculate the answer on the basis of the error, and with the possibility that valid marks will be missed. It also tends to complicate the presentation of the mark schemes, which need to provide the expected method and answer, any information about other acceptable answers, any information about what is not acceptable and any information about the degree of latitude for the follow through marks. Nevertheless, in terms of reliability, mathematics marking is extremely accurate.

MFL has a somewhat different relationship with accuracy and quality of expression from other subjects. Partly to avoid this, assessment of the receptive skills avoids requiring the students to produce the language. But for the expressive skills, marking covers content, quality of expression and accuracy. As noted above, quality of expression can place a very unhelpful hurdle in the way of the assessment. Creating a mark scheme that appropriately rewards accuracy is also problematic in that one student may produce a very safe response that is, as a result, relatively free from error whereas another may write ambitiously, using complex constructions and sophisticated vocabulary, but almost inevitably making more errors. The wording of MFL mark schemes can also suffer from a failure to recognise that it is an assessment of a foreign language, the top band seeming to require performance in terms that would stretch a native speaker.

It was also pointed out that there is a real difference in what is expected from an extended answer in, say, English literature and one in, say, law. Some subjects call for evidence of personal response in an essay (although backed up by knowledge and understanding), whereas others are looking for a carefully reasoned

logical argument citing appropriate authorities leading to a conclusion which itself may not be open. It is important that the level descriptors for these very different forms of extended writing clearly reflect these very different expectations.

Another important subject-specific factor is that the command words do not necessarily mean the same thing across all subjects. For example, the distinction between a description and an explanation is rather less clear cut in some aspects of science than it is in, say, some humanities subjects. It is therefore for each subject to determine its own taxonomy of trigger words and then use them consistently. It would be helpful if there was a great deal more commonality than that.

Further thoughts

It has already been pointed out that it is important to use qualifiers in band descriptors carefully and consistently (although what these are may be somewhat subject-specific). It is also important that they work as a hierarchy in natural language, which probably makes the maximum number of bands about five. But levels-based mark schemes exist in both GCSE and A level, and those very constraints of natural language may mean that the same qualifiers need to be used in both levels of examination. However, they should not mean the same: a limited GCSE answer should be more limited than an A level one and a sophisticated GCSE response less sophisticated than an A level one. It is probable that this can be best clarified through exemplification, which means that it cannot be tackled within a single qualification, especially as examination materials are increasingly in the public domain. Even if exemplar materials for both qualifications are not published, it is important to ensure that they illustrate not only progression within a qualification but also progression across levels.

7 Summary of AlphaPlus Focus Group

There was general agreement that the main area where concerns about marking accuracy arise was in the marking of questions requiring extended responses, usually in writing, so as a result discussion centred on the marking of extended answers. The main points emerging were:

- Holistic, analytical or points based

  It is important to mention all three approaches, but to stress that the use of a purely points-based mark scheme is rarely, if ever, used in GCSE or A level for high tariff questions. They may be valid for very knowledge-dominated items (although these would tend to be very reductive and better as structured items) or for where the marking team is inexperienced or otherwise non-expert, since it is a more mechanical process. There is evidence that such schemes are less reliable once the number of marks exceeds 12 (see Bramley, 2008).

  There are competing arguments about the two levels-based approaches. There is evidence that the analytical approach is slightly more reliable than the holistic one. Also, an analytical scheme ensures in principle that the marks reflect the intended weighting to the different strands/AOs being assessed. It is also more laborious and therefore resource-demanding and vulnerable to a halo effect, where the marker fails to keep performance across the different strands separate in awarding the marks. Conversely, the holistic approach more closely mirrors the authentic way of reading extended writing and is quicker and less demanding of resources. However, it increases the cognitive load on the marker in trying to establish best fit over a wider range of skills. (It is also almost certainly no less vulnerable to halo effects, but they are less easy to detect.)

  There is also the possibility of a kind of hybrid scheme, where a certain number of marks are awarded for making particular points relevant to the topic, and the rest for a holistic judgement about

coherence, cogency etc. This may be suitable for particular types of extended answer, but it would be important to ensure that the structure of marks properly reflected the balance of AOs in the construct.

- Role of comparative judgement

  We discussed the place of comparative judgement in the marking process, although it was acknowledged that this currently has limited applications in high stakes assessment. Comparative judgement remains an important potential approach to marking which would completely alter the nature of mark schemes. There are two aspects of comparative judgement as a process which it is important to recognise. The first is that it acknowledges that the process of marking and, in particular, marking extended answers, is essentially a series of comparisons, not only to the mark scheme but to a range of other models and archetypes. (Crisp's use of think-aloud protocols shows this clearly – cited in AlphaPlus, 2014.) The second is that it does not avoid the creation of a mark scheme. In fact, it could be argued that it requires two key elements of an effective mark scheme. At the very least, there needs to be an agreed 'importance statement' (Ahmed and Pollitt, 2011) trying to capture, in subject-specific terms, what the students are supposed to show about what they know, understand and can do. Such a process would almost certainly also benefit from the setter producing indicative content for an item, since it functions as an effective method of focussing on whether a particular task will produce the student behaviours expected.

- Number of levels

  The number of levels to be used as part of a mark scheme is partly a function of the total number of marks available, but it is also constrained by the limits of natural language. For example, when the decision was taken to increase the number of grades in GNVQ from three to four, a new term had to be added to the original Pass, Merit and Distinction. The solution, Credit, works as an intermediate term between Pass and Distinction, but it is not, in natural language terms, self-evidently inferior to Merit. As a guide to making a judgement, it would be little help, and that is what is required of the qualitative vocabulary in a mark scheme, which needs to be as transparent as possible. Of course, the extensive training that is the norm in achieving standardisation can flesh out the terms, but it leaves the process vulnerable, especially given the secondary audience of teachers and students who will not receive the training. A further problem (essentially created by the limits of natural language) is that the same terms are likely to be needed for examinations at different levels, as they were for GNVQ. But a term like 'limited' will be unlikely to mean the same thing in a GCSE mark scheme as in an A level one.

  In general, the more marks available for an item, the greater the level of inaccuracy (Bramley, 2008) and in particular the risk of bunching. Moreover, the bunching effect occurs even within a mark band if the band has a large number of marks available. It is another advantage of an analytical over a holistic approach that the total marks can be distributed over a greater number of cells in the grid, each with a specific descriptor, thus reducing the number of marks within a cell. However, this approach also produces some regression to the mean effects, where any psychological reluctance to award maximum marks is repeated within each strand of the assessment, thereby reducing the overall range of marks awarded.

- **Impact of online marking**

  There has been a significant shift towards online marking over the last ten years and this can have a significant impact on the design of mark schemes as well as, more obviously, on the processes involved. One area of significant benefit that online marking offers is that it greatly improves the level and efficiency of communication between marker and supervising marker. As a result, the process for dealing with unexpected answers is also improved, thereby reducing pressure on mark schemes to try to cover everything.

Furthermore, where the online marking system has markers marking a subset of questions, the mark scheme that each marker has to become familiar with automatically becomes more manageable and thus can become more extensive for each question without overload.

Conversely, it is not always possible to access the mark scheme on screen at the same time as viewing the question, making the mechanics of toggling either between screens (especially) or from screen to hard copy slightly more demanding cognitively. It is therefore important to think carefully about the user interface. This applies especially when dealing with extended answers that cover more than one screen, where the natural process of evaluative reading – checking back etc – is hindered.

There is also no longer such a clear distinction between marking and standardisation. There may be an initial and clearly defined co-ordination process, the standardisation meeting, with even this often carried out on-line. But the sampling stages are now generally built in to the marking process, using seeded items. This makes it much easier to re-focus (or stop) a marker who is mis-applying some part of a mark scheme.

- **How to deal with requirements to assess non-construct relevant elements**

  Essentially, all the expert advice on creating effective mark schemes stresses the need to ensure that the construct is clearly defined (c.f. the 'importance statement') and focusses on a single trait. If there is a high-level requirement to include some construct irrelevant feature (e.g. spelling, punctuation and grammar (SPaG)), it is best to assess it separately. In this case, the same principles of effective mark schemes apply to the sub-scheme: it is important to make sure the exact requirements are clear. Importantly, there will almost certainly be various halo effects. N.B. Where extended answers involve continuous prose, some elements such as making sure what you say is comprehensible (which almost certainly includes sufficient syntactic and grammatical control) are likely to be part of the construct and it is important not to reward these qualities twice.

- **Use of model answers**

  It is important to distinguish between exemplar responses (i.e. real answers drawn from pre-testing or live scripts) which, accompanied by annotations, are a useful mechanism for illustrating how to apply the mark scheme, and model answers (i.e. where the examiner writes a model of how the student is expected to respond). Model answers can be set at an unreasonably high standard and tend to make the highest marks seem unobtainable and thus reduce the effective mark range. They also can come to be interpreted as required material (which is always a risk when providing exemplification) and depress all the marks.

  There is a case for using exemplar responses as part of the published mark scheme, and they may be essential operationally if there is to be no training in the use of the mark scheme. But, even here, it is unwise to use model answers, since they give no idea of the true expectations of standards.

- **Role of mark scheme in preventing predictability**

  One of the regular complaints among some stakeholders is that exams have become predictable. There are several factors in terms of the specification and the process by which it is sampled over time which may contribute to this. For example, too obvious a rotation of topics may result in the taught content being less than the specification intends or expects. Or wording in items which is too vague or varies too little across exam series may result in pre-prepared answers being over-rewarded. These issues are outside the remit of this work. However, the use of standard, generic levels-based mark schemes, which are common to all questions on a paper and even from year-to-year, is often seen as a cause of predictability. There are two key remedies to this beyond setting better items. The first is to ensure that relevance is a key feature of the level descriptors. The second is through the indicative content. It

is here that the demands of each specific question reside, so it is important that it captures those demands clearly rather than being over-general.

- **Impact of availability of marked scripts and marking schemes**

  It is important to recognise that the return of marked scripts and the publication of mark schemes are now features of the system. This has played a major role in improving the transparency of examinations to teachers, students and the general public but it brings with it extra challenges. Given that this is a secondary audience, it must not be allowed to interfere with the main purpose of a mark scheme, making the marking as accurate as possible. In an important sense, it may be better to see the scheme as being educative for teachers: showing them the issues that arise with marking in deciding, for example, what is an acceptable answer even when it is not absolutely correct. (Where the teaching force may be relatively inexperienced in the subject, as is sometimes the case, this educative purpose is especially important.)

8 Discussion

There is extensive common ground between the findings of the three strands of evidence collection. All are agreed that the question and mark scheme must be developed in tandem. In addition, there is widespread endorsement of the need for the mark scheme to identify the principle on which credit should be awarded rather than try to provide an exhaustive list of acceptable and unacceptable answers.

Moreover, practice is clearly very similar across awarding bodies, with papers containing similar item types (at least within a subject) which are marked in much the same way whoever has set them.

There is general agreement about how to use exemplar material and, in particular, its part in standardisation meetings where it has a key role in helping the markers understand how to interpret the language in the mark scheme. Indicative content is seen as important for helping to define how the item is addressing the defined part of the specification. However, although it is important, everyone recognised the need to make sure that it was kept separate from the part of the mark scheme which showed how to award credit for the students' responses. In a small number of cases, particularly where a medium tariff question was addressing a single AO, it was suggested that some elements of the indicative content could usefully focus the level descriptors.

One point which emerged with great consistency in the expert interviews and which was also picked up in the literature review is that the publication of mark schemes means that they have an important secondary audience. This presents particular challenges, especially if there are permitted responses in the mark scheme which are likely to be controversial. There was, however, a clear recognition that such considerations must not be allowed to interfere with the mark scheme's main purpose as the primary channel of communication between the setter and the marker. In fact, it was suggested that the existence of such decisions could be useful in building understanding between awarding bodies and other stakeholders, for instance in awarding body organised continuous professional development meetings with teachers.

One point that received universal recognition was the need to balance the thoroughness of the mark scheme with making it manageable. In many cases, the inherent complexity of the process of marking means that the mark schemes are close to the limit in terms of cognitive demand. Several of the points raised seek to address this problem. First, there is the idea that, for each item, the scheme should start with a clear statement of the qualities expected in a response, usually in terms of the AOs in the specification. This constitutes a principle for how to award credit and means that there is less need to try to provide exhaustive lists of what are or are not acceptable answers.

Second, it was pointed out that a lot of the bulk of mark schemes comes from repetition of the same piece of guidance. This suggests that there are also higher order principles about marking, possibly subject-specific or item-type-specific. It may, therefore, helpfully reduce the cognitive load if as many of these principles are

extracted from the body of the mark scheme and presented as a separate section at the beginning. Of course, this carries the risk that such general principles may be forgotten if they are not present within the marking scheme for an item. For example, a particular danger highlighted in the literature review and emphasised throughout the expert interviews is the need to remember that indicative content does not constitute required material for an answer (although clearly some pieces of knowledge will be essential for dealing with tasks in a science paper). Where it is apposite to take this into account, it probably bears repeating.

This would accord well with the suggestion made in several of the interviews that the best way to help with interpretation of the qualifying adjectives in level descriptors would be through a set of exemplars, with commentaries, perhaps drawn from previous years' responses – a document, it was noted, that would be very useful in terms of relationships with schools. It should be noted that this would only be possible if there were good reason to use the same levels-based structure across several items and series.

The language used, particularly in level descriptors, raised another issue. Not only do the terms used need clarification, but they also need to carry the meaning they do in natural language. This places some constraints on the number of levels that can be used. It is important, for example, that the top band is accessible to the best students by avoiding absolute terms or, as was pointed out, being as good as a native speaker in MFL. (Of course, it is also important in terms of public perception that the demands of the top band are not made too modest.) Similarly, it is important that the bottom band, outside a mark of 0, is seen to be based on some real attainment, rather than just the absence of qualities that would earn higher marks.

There were consistent concerns about the presence of hurdles (helpfully backed up by research by Ofqual). These may be deliberate (reflecting the idea that a student who doesn't display a particular skill should not get into a particular mark band) or almost unintentional, where the attempt to clarify the qualities usually displayed by a good answer fail to recognise that this does not apply to all answers generally recognised as good.

In the end, the level of consistency both within and across the various sources explored means that it is possible to extract a set of general principles and guidance. In no particular order, these are:

- There must be alignment between the specification, the item and the mark scheme, with the mark scheme clearly showing the feature(s) of the specification being tested. During the development process, if a change is suggested, maintaining the mark scheme's relationship to the specification takes precedence.
- By extension, weightings of different elements in a mark scheme should be explicit. An analytical mark scheme is more helpful in this respect than a holistic one. In this regard, the marks available should reflect the demand of the question.
- By further extension, it is important to ensure that the command words used in a question align with the expectations of the mark scheme. It is only if the wording of the question controls as far as possible the behaviours of the student that the mark scheme can fairly reward those behaviours. This includes issues such as the distinction between describe and explain, but also covers requirements such as being explicit about the number of examples required, if this will affect the mark.
- It is important to get the balance right between the level of detail provided and not overloading the examiner. This applies whether it is a one-mark item or a 25-mark essay. If during development, a mark scheme is becoming too long or too complex, it may be necessary to revisit the question to further limit the range of possible responses. The clear statement of the principle underlying the award of credit should reduce the need for extensive exemplification.
- There should not be hurdles built in to the mark scheme. There is very strong evidence that these distort the rank order and reduce the accuracy of marking. If it is felt that a particular skill or piece of knowledge should be valorised over others, this should be reflected in the structure of the specification. Beware of unintended hurdles!

- Model answers do not help. Their main effect is to depress the marks awarded to good responses, and they do little to illustrate the expectations of the mark scheme for real student responses. (Clearly, for something like a mathematics problem, there will be at least one fully worked solution in the mark scheme, but the importance of these is primarily to illustrate how the marks are awarded. Similarly, an item requiring a definition may well provide the dictionary definition as part of the mark scheme, but the main task of the mark scheme is to help the markers decide how to reward incomplete or imperfect definitions that they will encounter).
- Exemplification should be about the specifics of the language or mark scheme requirements. As noted above, this exemplification may be better as a separate document which can be applied to all questions of a given type in a paper or even specification. It would be the task of the training and standardisation process to apply the information in such a document to the specific questions in the current assessment.
- It is vital for the setter to have a clear, shared and preferably explicit view of the construct being assessed. This may be implicitly captured within a statement of aims in the specification.
- It is important to put a lot of effort into driving home the idea that there is no required answer to most higher tariff items. Some of the presentational features sometimes used in mark schemes, such as bullet points or the use of emphasis, may actually make this harder, while ordering of mark bands from low to high appears to help.
- Indicative content is important but should be kept as distinct as possible from the mechanics of how credit is to be awarded. Rather, indicative content serves as another mechanism for capturing what the setter had in mind when setting the question. It is useful, therefore, in checking that the question, mark scheme and specification are remaining in line and that the indicative content truly reflects the sum of the content which the particular question is going to evoke in a student.
- The publication of mark schemes may well be the single greatest contribution to the examination system. However, it does mean they have a significant secondary audience – they are being used as an important tool for teaching and learning. This cannot be ignored, but it should not deflect them from their primary objective of ensuring marker consistency. Rather, it may be better to see the availability of mark schemes as an important element in improving stakeholders' understanding of assessment.
- Precisely because a mark scheme doesn't work if it is too long or complex, there is little point in aiming for comprehensive coverage of what is acceptable or not. Indeed, as noted above, it is a better strategy to clarify the principle behind the award of credit, the approach that markers are expected to adopt. Importantly, too, the mark scheme is not seen in isolation but as an, admittedly central, part of the overall standardisation process. Training scripts, standardisation scripts and discussions at a meeting also have a major role in helping markers understand what is required. (The move in some cases using on-line marking to limit each marker to a subset of the questions should also make it easier to co-ordinate the marking.)
- On the whole, as many unexpected answers as possible are captured during the overall standardisation process. Otherwise, they are mainly handled through referring to a more senior marker – a process which has been greatly enhanced and streamlined by the advent of on-line marking. However, it would be helpful to provide some guidance as to how to approach the assessment of a response outside the mark scheme. After all, the mark scheme needs to be an empowering document.
- A particular factor to consider with answers which don't match the mark scheme is the use of pre-prepared answers or some other feature that has made the question too predictable. The mark scheme needs to be constructed to mitigate this risk.
- On-line marking delivers many benefits in terms of marker accuracy. However, it brings some disadvantages especially in the assessment of extended answers. The marker cannot place the mark scheme and response side-by-side as they can on paper; it is harder to toggle across screens if a response goes over more than one screen as is natural when reading. And it is harder to go back, in the case of second thoughts which are an inevitable element of making highly complex judgements. At the least, the mark schemes (and marking platform) should be designed to minimise these factors.

- On-line marking also makes it more-or-less impossible to detect some forms of malpractice and some features which are relevant to fair and accurate marking. In subjects with predominantly short-answer questions, an alert marker could detect collaboration through identical wrong answers; with extended answers, especially those requiring personal response, it is often useful to see all the answers from a centre to judge whether a response is displaying such a quality.

9 Implications for Mark Scheme Guide

The principal deliverable for Qualifications Wales from this project is the production of a practical guide to mark scheme design. The need for further guidance was supported by the conclusions of the literature review and therefore, in the course of the expert interviews, we raised this as a possibility and also canvassed ideas and opinions about what should be included in a mark scheme guide. Below are listed some of the points made:

- There was support for the idea of further guidance and an expert system
- It would need to be clear and succinct: like a mark scheme it would need to get the balance right between thoroughness and overload
- It would need to provide examples, both good and poor
- It would need to make clear the strengths and weaknesses of each possible approach
- It would be very useful to provide checklists of the kind of actions needed when reviewing a mark scheme
- Flow charts and diagrams would be helpful to illustrate key processes, but there is also the need for full explanations of concepts
- It would need to reflect differences across item types and across subjects or subject areas
- It would have to consider the need for, and limitations of, qualitative vocabulary in differentiating between descriptors
- It would have to help with understanding the development process as well as evaluating the outcomes.

These concepts represent a useful starting point for any subsequent guidance.

10 Appendix 1: Interview Schedule

Effective mark schemes: interview schedule

Introduction

- Introduce yourself and Alpha*Plus.*
- Explain the nature of the project and the purpose of the interview.
- Explain that the interview should not take more than 40 min unless he/she (the interviewee) is happy to go a bit over. [Interviewer to try to pace the interview to get through in 40min. If the interviewee has a lot to say, check at around 30 minutes, whether happy to go a bit over.]
- Check whether they are happy for the organisation they represent and/or their role to be listed in the report we submit, and whether they would be happy to have their views attributed
- Ask whether the person is happy for the interview to be recorded for use by the AlphaPlus research team only.

If yes, start voice recorder. If not, note the essence of responses on paper or as a typed document during the interview.

Set parameters of the discussion

We are interested in a range of issues:
Principles of what makes a good mark scheme
What they would like to see in a mark scheme guide
different styles of mark scheme
the range of constructed response item types to be covered
the expected level of (mark/remark) reliability

A. Find out more about the interviewee

Please can you start by telling me a little more about your background and experience in relation to writing/evaluating mark schemes

For example, have you focussed on particular subjects or modes of assessment
What exactly has your role been e.g. development, evaluation, administration etc

B. Writing mark schemes for short answer items

What are the best approaches for short answer items?
Have you seen examples where this has been done well? (Can you share examples?)
How do you deal with unexpected answers?
How do you provide appropriate exemplification?
What should be the role of indicative content?
Where does it generally go wrong?

C. Writing mark schemes for items requiring extended answers

What are the best approaches for dealing with extended writing?
Have you seen examples where this has been done well? (Can you share examples?)
How do you deal with unexpected answers?
How do you provide appropriate exemplification?
What should be the role of indicative content?
Where does it generally go wrong?

D. Subject differences

Are there subject-specific factors that affect the design of mark schemes?

Can you give any examples?

Further thoughts

What do you think would be useful in a mark scheme guide?
Is there anyone else you think we should speak to?
Is there anything else you would like to tell me about the topic?
Thank you very much taking the time to talk to me today, etc…

11 Appendix 2: Literature Review References

Ahmed, A. and Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes, Assessment in Education: Principles, Policy & Practice, 18:3, 259-278.

AlphaPlus (2014). Standardisation methods, mark schemes, and their impact on marking reliability. Ofqual/14/5380.

Black, B., Suto, I. and Bramley, T. (2011). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement, Assessment in Education: Principles, Policy & Practice, 18:3, 295-318.

Bramley, T. (2008). Mark scheme features associated with different levels of marker agreement, Paper presented at the British Educational Research Association (BERA) annual conference, Heriot-Watt University, Edinburgh, September 2008.

Child, S., Munro, J. and Benton, T. (2015). An experimental investigation of the effects of mark scheme features on marking reliability. ARD Research Division Cambridge Assessment.

Christie, M., Grainger, P., Dahlgren, R., Call, K., Heck, D. and Simonet, S. (2015). Improving the Quality of Assessment Grading Tools in Master of Education Courses: A Comparative Case Study in the Scholarship of Teaching and Learning. Journal of the Scholarship of Teaching and Learning, Vol. 15, No. 5, October, 2015, pp.22-35.

Ghalib, T. and Hattami, A. (2015). Holistic versus Analytic Evaluation of EFL Writing: A Case Study, English Language Teaching 8 (7), 225-236.

Lille, B. and Romero, M. (2017). Creativity Assessment in the Context of Maker-based Projects, Design and Technology Education: an International Journal, 22 (3).

Ofqual (2014). Review of Quality of Marking in Exams in A Levels, GCSEs and Other Academic Qualifications, Ofqual/14/5379.

Pinot de Moira, A. (2011). Levels-based mark schemes and marking bias, CERP, AQA.

Pinot de Moira, A. (2013). Features of a levels-based mark scheme and their effect on marking reliability: Centre for Education Research and Policy, AQA.

Tisi, J., Whitehouse, G., Maughan S. and Burdett, N. (2013). A Review of Literature on Marking Reliability Research (Report for Ofqual), Slough: NFER.

12 Appendix 3: Literature Review Specification

Literature Review Specification

This document sets out the requirements of the literature review to support the development of the mark scheme guide for Qualifications Wales.

Overall brief

This project is to develop a mark scheme guide for use by awarding bodies in Wales. The development of this guide is to be supported by a literature review and stakeholder interviews.
The project, therefore, has three aspects:

- A review of the academic literature
- Stakeholder interviews
- Producing the Mark Scheme Guide

The literature review and interviews are supplementary, to inform the development of the main output – the mark scheme guide.
The overall question the literature review aims to address is:
Primary question:
What does the academic literature tell us about the principles of good mark scheme design?
Secondary question:
What features of the mark scheme support accurate marking and valid assessment?

Search terms

Based on what we know about the requirements of this review, we propose including the following terms as part of the search design:

- 'Mark schemes' + reliability
- 'Mark schemes' + validity
- 'Knowledge assessment' + 'mark schemes'
- Exam + 'mark schemes'
- 'Exam mark schemes'
- Good + 'mark scheme design'
- 'Mark scheme' + design
- 'Scoring rubric'
- 'Effective' + 'scoring rubric'
- 'Scoring rubric' + 'design'

We will include literature from 2000, to include any jurisdiction, but published in English.
Where possible, research literature will focus on the assessment of 14 to 19 year-olds either in general or vocational education.
If high quality literature reviews are found from recent publication, we will use that as a basis and supplement this with later literature.
A Google search was also undertaken. The terms used include

- 'what makes a good mark scheme'
- 'mark scheme design'

Methodology

We will undertake the following stages:

- Identify suitable sources/database, and establish/refine relevant search terms
- Undertake the literature search and produce a 'long list' of relevant references
- Review the long list and refine this down into a short-list of approximately 12 good quality sources
- Summarise the literature using the agreed template
- Use the summaries to write-up the literature review.

The following websites/ organisations are identified as potentially interesting and will be added to/ used at the initial search stage:

- WJEC, Edexcel/ Pearson, OCR/ Cambridge Assessment, AQA, City and Guilds

Outputs

The following outputs will be produced:

1. Long list of sources (references) identified
2. Short list of 12 most relevant sources, and where possible the full text or a link to the full text of each
3. A written-up literature review in Word format. There is no word limit, but concise and with key findings / recommendations pulled out.