A Framework for the Qualitative Analysis of Examinee Responses to Improve Marking Reliability and Item and Mark Scheme Validity

Ezekiel Sweiry
Standards and Testing Agency

The predominant view of validity is based on the Messick (1989) argument that validity is not a property of the test, but rather a property of the meaning of the test scores. Some researchers caution that this view risks ignoring the vital role of question and mark scheme developers in ensuring that assessments are valid. Pollitt et al. (2008), for example, propose a different conception of validity, which requires that the cognitive processes elicited by the question are those intended by the question writer. Validity in this sense can only be investigated through a thorough exploration of the student thinking that is triggered by a question.

Regardless of the conception of validity employed, a test cannot be valid if it does not produce scores that are consistent and relatively free from error. Reliability is therefore a necessary condition for validity, and marking reliability represents the greatest threat to the reliability of many assessments that use constructed-response items.

This paper presents a framework for using qualitative evidence captured from item responses to improve item and mark scheme validity and marking reliability in constructed-response items.

[Key words: validity, marking reliability, construct irrelevant variance]

A paper for the 39th Annual Conference of the International Association for Educational Assessment, Tel Aviv, October 2013

zeek.sweiry@education.gsi.gov.uk
Standards and Testing Agency
Sanctuary Buildings
Great Smith Street
London SW1P 3BT

Introduction

The predominant view of validity is based on the Messick (1989) argument that validity is not a property of the test, but rather is a property of the meaning of the test scores. This view holds that an assessment will only be valid if those who use its results make appropriate interpretations of those results. However, some researchers caution that this conception of validity risks ignoring the vital role of question and mark scheme developers in ensuring that assessments are valid. More specifically, Pollitt (2009) states that there is a danger that with this view, validity will be seen as 'the business of test interpreters, *rather than* of test constructors'. Pollitt el al. (2008) propose an alternative view of validity, arguing that 'a test question can only contribute to valid assessment if the students' minds are doing the things we want them to show us they can do; and if we give credit for, and only for, evidence that shows us they can do it'. In this view, the essence of validity is the interaction between a test question and the student's mind: if this goes wrong then validity is impossible. It is this alternative conception of validity that is employed throughout this document.

Regardless of the conception of validity employed, a test cannot be valid if it does not produce scores that are consistent and relatively free from error. Reliability is therefore a necessary condition for validity, and marking reliability represents the greatest threat to the reliability of many assessments that use constructed-response items.

Validity can most practically be investigated, and ultimately improved, through a qualitative analysis of the responses students give to test items that is designed to establish whether the cognitive processes elicited by the question are those intended by the question writer. In addition, a qualitative analysis of responses is also a highly effective means of investigating and improving the marking reliability of items that use analytical (points-based) mark schemes. This paper considers how a qualitative analysis of the responses to test items and the scores awarded to those responses can be used to improve item and mark scheme validity and marking reliability.

The techniques proposed in the paper can be most effectively employed in assessments where items are pretested before appearing in a live test. Where this is not the case, the methods focused on improving marking reliability can still be used if it is possible to carry out a qualitative analysis of responses before marking takes place. In addition, a post-test qualitative analysis of responses can be a powerful way of improving the validity of future test items. The key to writing good items is to understand how students think when answering test items, and the qualitative analyses proposed in this paper will help question writers to anticipate how students will react to the test questions they write.

Response Analysis

The term *response analysis* is defined here as the systematic qualitative analysis of responses to a question (and, where applicable, the score awarded by a marker to those responses), and the active consideration of whether, for each response, there is evidence that:

- The mark scheme does not provide sufficient guidance to markers for it to be marked reliably.
- Students are either losing or gaining credit for construct irrelevant reasons.
- The mark scheme, when applied correctly, does not result in appropriate credit being awarded.

Standards & Testing Agency

Response analysis and marking reliability

The impact of scoring on validity is typically described in the context of reliability, and is measured in a variety of ways, including the extent of agreement between markers and 'definitive' scores set by experts, and in terms of levels of consistency both across and within markers. A response analysis that focuses on the marks awarded to responses by markers (when items are pretested) is an effective means of improving the marking reliability of items that use analytical (points-based) mark schemes. For any item using an analytical mark scheme for which it is not possible to list, word-for-word, every creditworthy response in mark scheme, there is inevitably a degree of subjectivity involved in the marking of some responses. The two questions below use the well-known children's story 'Jack and the Beanstalk' to exemplify this point. In item 1, the mark scheme is highly constrained, and it is possible to list every creditworthy response in the mark scheme. (In fact, there are only two: 'milking cow' and 'cow'.)

Item 1

Text

Once upon a time there was a poor widow who lived with her son Jack in a little house. Their wealth consisted solely of a milking cow. When the cow had grown too old, the mother sent Jack to sell it.

Question

What did Jack's mother decide to sell? (1 mark)

Mark scheme

Award 1 mark for the answer: Milking cow **or** Cow

At first glance, item 2 appears to be another straightforward-to-mark short constructed response item. However, in item 2, it is *not* possible to list every possible creditworthy response in the mark scheme. Instead, markers are required to evaluate responses against the criteria stated in the mark scheme. For marking to be reliable, the criteria (and associated guidance such as example responses) must provide sufficient information for markers to do this accurately and consistently. Even in such a straightforward question, however, this is more difficult to achieve than might be expected.

Item 2

Text

Then, at the height of her exasperation, Jack's mother threw the five beans out of the window and sent Jack to bed with no dinner. The morning after, when he stepped outside, Jack saw an amazing sight. A giant beanstalk, reaching far into the clouds, had grown overnight. "The beans must have really been magic," Jack thought happily.

Question

How did Jack know that the beans were magic? (1 mark)

Mark scheme

Award 1 mark for reference to the fact that a giant beanstalk had grown **and** that the beanstalk had grown overnight , eg A huge beanstalk had grown since yesterday

For example, if the response refers to a 'beanstalk', rather than a 'giant beanstalk', would this be creditworthy? And if this response is creditworthy, would 'plant' still be worthy of a mark? If not, what about 'giant plant'? Or 'tree'? And if, rather than state that it had 'grown overnight', a student responds that a giant beanstalk 'had grown suddenly', would this be worthy of the mark? Ultimately, the minimally acceptable expression of each of the two elements of the marking criteria (that a giant beanstalk had grown and that the beanstalk had grown overnight) is not sufficiently defined in the mark scheme.

The potential for marking disagreement in the two items can be better understood through the concept of the zone of uncertainty (ZoU). The ZoU is defined as the range of responses to an item (actual and potential) for which it is unclear whether the mark scheme criteria are satisfied. A qualitative analysis of responses to constructed-response English reading, maths and science test questions that used analytical mark schemes, carried out by Sweiry (2012), found that the greatest threat to marker agreement levels was the size of the ZoU.

The mark scheme for item 1 has no ZoU, i.e. there are no responses (actual or potential) for which it is unclear whether the mark scheme criteria are satisfied. Figure 1 shows that the range of actual and potential responses can be categorised into discrete sets through application of the marking criteria. The mark scheme for item 2 has a sizeable ZoU, shown by the overlap between the ranges of responses worth 0 marks and 1 mark in figure 2.

Figure 1: Zone of Uncertainty for item1

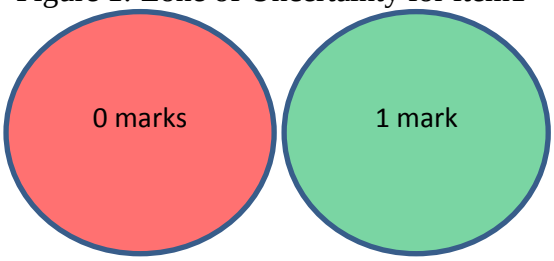
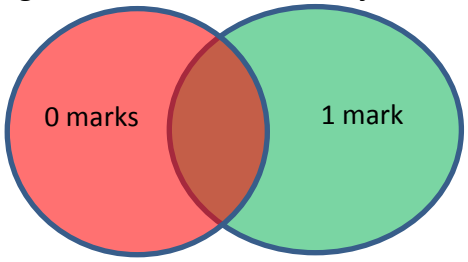


Figure 2: Zone of Uncertainty for item 2



Response analysis can be used to identify responses that markers found most difficult to score or tended to score inconsistently. (Alternatively, if the responses have not yet been marked, response analysis can be used to identify responses that are judged likely to be difficult to score using the mark scheme criteria.) The evidence collected from response analysis can then be used to reduce the ZoU. One method of minimising the ZoU is through the effective utilisation of difficult-to-score responses within the mark scheme. Effective utilisation requires that:

- Both creditworthy responses *and* non-creditworthy responses are recorded and used within the mark scheme.
- 'Marginal' non-creditworthy responses (i.e. those that are only just insufficient to receive credit) and creditworthy responses (those that are only just sufficient to receive credit) are uncovered through the response analysis and used within the mark scheme.

A more comprehensive approach to reducing the ZoU is through the development and use of themed response tables (TRTs). TRTs are tables, populated with creditworthy and non-creditworthy responses that were given to a question and identified (through response analysis) as being difficult to mark. The different types of responses that could be given to the question are separated into 'themes', with each row in the table corresponding to a

different theme. For each theme, the '0 marks' column shows examples of the *best* responses that are just insufficient to be awarded a mark, while the '1 mark' column shows minimally acceptable responses (i.e. the lowest quality responses that are just sufficient to be awarded the mark). The tables are then provided to markers, who are encouraged to refer them in situations where the mark scheme does not provide sufficient guidance.

TRTs are appropriate for all items that use analytical (points-based) mark schemes. For multi-mark items, additional columns can be added for the additional mark points, or rows could be used to represent marking criteria (or creditworthy points) in the mark scheme rather than 'themes'. By providing these tables in addition to the mark scheme rather than as part of it, the additional cognitive load placed on markers is minimised, as markers only need to refer to the tables for responses where the mark scheme does not provide sufficient marking guidance.

In a typical mark scheme, even where correct and incorrect responses are shown, it is difficult for markers to assimilate all relevant information effectively to form an understanding of where the 'cut-offs' lie between correct and incorrect responses, because related correct and incorrect responses are not clearly linked. The horizontal (row-based) presentation used in TRTs allows related correct and incorrect responses to be shown adjacent to each other.

Table 1 below shows a TRT for item 2. It can be seen that additional notation, in the form of underlining and square brackets, has been used. The former is used to indicate the specific element of a response that makes it creditworthy, while the latter is used to provide a rationale for the score given to a particular response.

Table 1: Themed response table for item 2

Theme	**0 marks**	**1 mark**
Size of the beanstalk	A big beanstalk had sprouted overnight A plant/ big plant/large plant/ tall plant had grown overnight The beanstalk had grown in one day	A <u>giant/huge</u> beanstalk had grown overnight A <u>huge</u> plant/ <u>giant</u> plant had grown overnight A <u>tree</u> appeared overnight The beanstalk went <u>to the clouds</u> in one day *[Accept 'to the clouds' as a sufficient reference to the size of the beanstalk]*
The speed at which the beanstalk had grown	There was a giant beanstalk The beans grew into a huge plant A huge plant was there	<u>Suddenly</u> there was a giant beanstalk The beans grew into a huge plant <u>overnight/ in a day</u> He saw how <u>quickly</u> the plant had grown and how big it was

Following a successful pilot, TRTs are now used to support the marking of the UK national assessments in English reading for 11 year olds. Feedback was gathered from the team leaders (leaders of the individual marking teams) involved in the marking of the 2013 test, and all 88 respondents agreed or strongly agreed that the TRTs were effective in improving the marking consistency of 'borderline' responses and increasing marker confidence.

Response analysis and item validity

Item validity requires that the cognitive processes elicited by the question are those intended by the question writer, and a response analysis is an effective means of investigating whether this is the case. A useful theoretical framework for considering the range of responses that can be given to items, and the implications for validity, is *outcome space.* Pollitt et al. (2008) define a question's outcomes space as 'the set of all responses to it, both actual and potential'. They classify the responses that can be given to a question into six categories, depending on whether they are anticipated or not anticipated, observed or not observed, and good or bad responses to the question. A high degree of overlap between observed and anticipated responses (both good and bad) to a question is desirable, as it would suggest that students were engaged in the type of thinking intended by the question setter.

A poor degree of overlap suggests that a question is not working as intended, and the most powerful indication of this are unanticipated responses that are seen more than once. Unanticipated responses represent outcomes that were not expected by the question writer and are therefore not reflected in the mark scheme. When unanticipated responses are seen, the cause of the response should be established. Incorrect unanticipated responses are of particular concern, as they may indicate that a question was not interpreted as intended. When the same unanticipated response is seen more than once, it is important to determine whether it appears to be the result of a lack of relevant subject knowledge, or if a feature of the question interfered with the students' thinking and triggered the response. Clearly, the latter would represent a threat to the validity of the item.

Some of the threats to item validity that are most likely to elicit unanticipated responses are summarised in table 2. It is important to state that this is not an exhaustive list. In addition, different assessments (based on subject, question format and other relevant variables) are likely to possess their own set of common associated threats.

Table 2: Common causes of incorrect unanticipated responses to test items

Overlooking of key word, phrase or other question element	Students may overlook a key word or phrase in a question, and as a result develop a different understanding of the question to the one intended. This can happen for a number of reasons: o If a question (including any associated visual resources and contextual information) contains more than around 5 or 6 ideas, it is possible that one or more of these will not be internalised by students due to the excessive strain on working memory. o Diagrams tend to dominate students' mental representations of questions at the expense of text (Crisp and Sweiry, 2006). o Students tend to read what they expect to, based on past experience. This is normally an efficient strategy, but can fail if there are aspects of questions that run counter to these expectations (Crisp et al. 2008).

Ambiguous language	Particular words or phrases within a question may have more than one meaning, leading to a different interpretation of the item than intended.
Inappropriate or vague command words	The command word (eg explain, describe, how) used in an item may be inappropriate given the type of response the item was intended to elicit. For example, the word 'explain' is often used when only a simple statement or description is required. In addition, the meaning of command words within questions is not always clear, and indeed many of them do not have universally agreed definitions (Pollitt et al. 2008).
Misuse of Context	The contexts used in the questions can lead to wrong ideas being activated in a number of ways: ○ Some contexts may be too complex, abstract or novel for students (Pollitt and Ahmed 2000). ○ When contexts are very familiar to students, students may answer questions based on their everyday knowledge of a context when what is actually required is an answer based on content within the content domain (Pollitt and Ahmed 2000). ○ There may be an expectation on the part of students that, where a question follows a detailed context, the answer is related to the information contained within the context, when in fact a simple 'textbook' answer is required (Pollitt and Ahmed 2000).
Inappropriate answer space	The size and format of the answer space in a question may present problems if it is not consistent with the type or length of answer that is required in a successful response. For example, providing a full line of answer space when only a one-word answer is required may cause students to re-assess what they think is expected by the question. A similar effect may result if too little space is given.
Item interaction effects	The response a student gives to an item may be affected by other items (and particularly immediately preceding items) in the test.

Correct unanticipated responses can also indicate problems with a test item or its mark scheme. In some cases, the mark scheme may simply be incomplete, and the problem remedied through basic additions to it. In other cases, however, unanticipated correct answers can signal more fundamental problems with a test item or mark scheme. For example, the potential range of correct answers to the question may be very broad, making the question much less demanding than anticipated.

Response analysis and mark scheme validity

Pollitt et al. (2008) state that 'it is not enough to write good exam questions that ensure the students' minds are doing the things we want them to show us they can do; our validity principle demands that we also give credit to, and only to, the evidence that they can do these things'. An analysis of responses, alongside an analysis of the scores given to the responses, can be effective in identifying mark schemes (or elements within them) that lead to scoring that does not meet this principle.

Many mark scheme issues can be identified through a 'credit discrepancy'. This means that the credit allocated to a response when the mark scheme is correctly applied appears to be excessively harsh or lenient. Some of the most common causes of credit discrepancy are

summarised in table 3, although it is again important to state that this is not an exhaustive list, and as with item validity, different assessments are likely to have their own set of common associated threats.

Table 3: Common causes of credit discrepancy

Mismatch	This threat refers to a mismatch between the task set by the question and the way in which credit is awarded (Pollitt et al. 2008).
Credit for unlikely content	In some cases, a high score on a multi-mark item requires content that students are unlikely to include, either because it is beyond the ability of the test takers or because students felt the content to be too obvious to state.
Poor differentiation	In multi-mark items, correct application of a valid mark scheme should lead to more marks being awarded to students who show evidence of greater achievement.
Inappropriate thresholds between mark points	In one-mark items in particular, marking criteria are often initially set either too leniently or stringently. A response analysis is an effective means of uncovering this.

In some cases, the source of any validity issue can be located specifically within the question or the mark scheme. For example, ambiguous language (table 2) refers specifically to an ambiguity in the task set by the question. However, in other cases, validity issues are caused by the way in which the item and mark scheme relate. One such example is mismatch (table 3), which refers to a discrepancy between what the task requires students to do, and the way in which credit is awarded. Clearly, in cases such as this, the source of the validity issue could conceivably be either the item or the mark scheme. Because the relationship between item and mark scheme may be central to validity, it is essential that the item and mark scheme are considered together rather than as separate entities when investigating threats to validity.

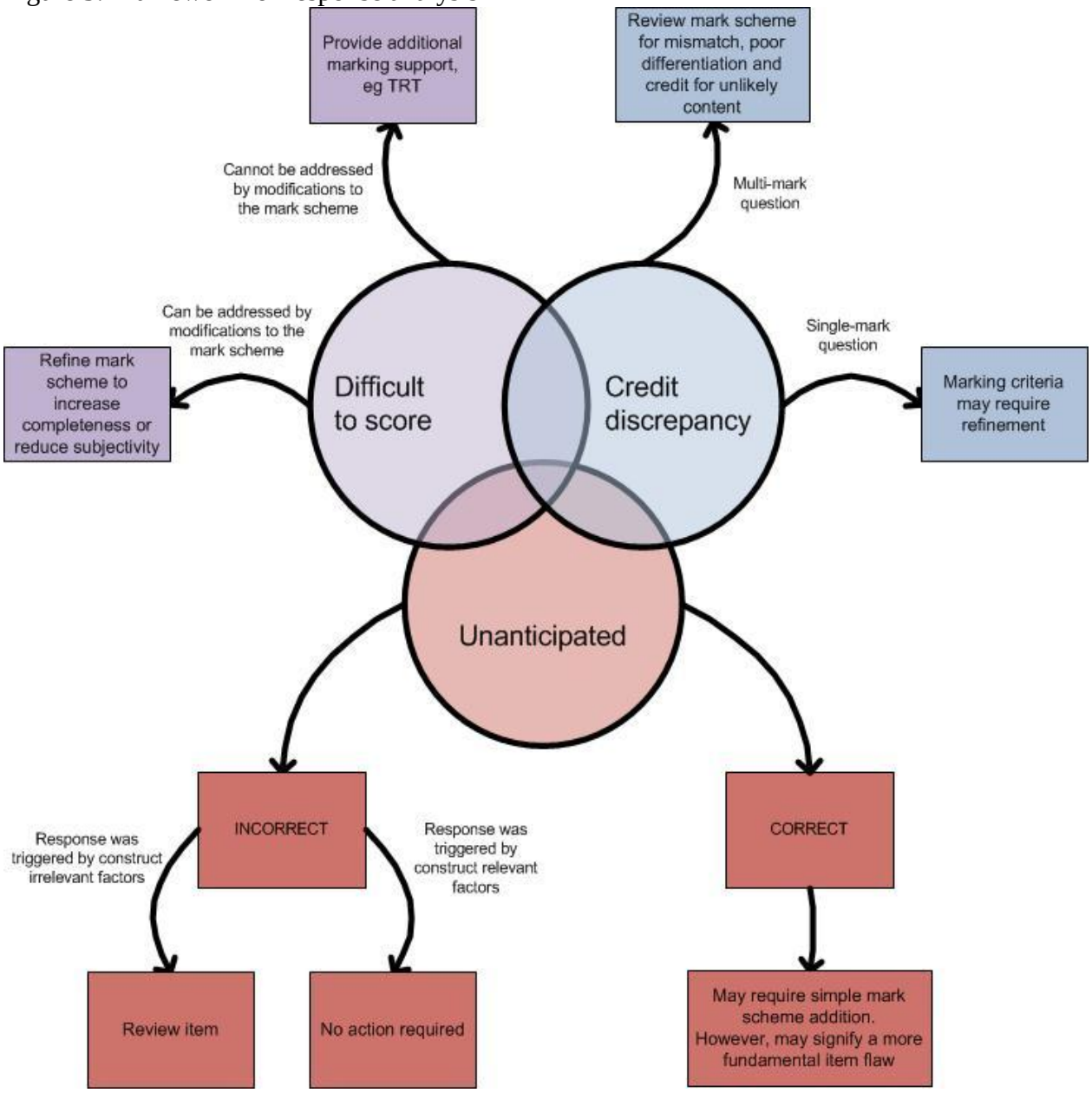
A framework for response analysis

The final section proposes the basis of a practical framework for the qualitative analysis of responses to improve marking reliability, and item and mark scheme validity, based on the methods and theory included in the previous sections of the paper.

The first stage is a response analysis that focuses on the responses given to test items and, where applicable, the scores given to those responses by markers. As shown in figure 3 (overleaf), responses should be recorded when they fall into one or more of the following categories: *difficult to score*, *unanticipated*, and *credit discrepancy*.

The centre of the diagram shows the three categories against which it is proposed that responses are recorded. The overlap between the categories shows that a response could highlight multiple validity issues and therefore may need to be recorded against more than one category.

Figure 3: Framework for response analysis



Responses recorded as *difficult to score* can be used to reduce the ZoU and ultimately improve marking reliability. In some cases this can be achieved simply through refinements to the mark scheme. In other cases, however, the amount of information that would need to be added to the mark scheme will likely compromise the accessibility of the mark scheme. In this situation, additional marking guidance may be required. This paper has outlined how themed response tables may be a highly effective means of reducing the ZoU in analytical (points-based) mark schemes.

Responses recorded as *unanticipated*, when seen more than once, may be an indication that an item has an underlying validity issue. Incorrect unanticipated responses should be analysed carefully to establish whether they have been caused by construct irrelevant

elements within the item, such as ambiguous language, excessive language or misuse of context (see table 2). Correct unanticipated responses can be caused in a variety of ways. In some cases the problem can be fixed simply through mark scheme refinement. However, in other cases such responses may indicate a more fundamental problem with the test item that cannot be solved through amendments to the mark scheme.

Responses should be recorded as showing a *credit discrepancy* when the credit allocated to the response, when the mark scheme is correctly applied, appears to be excessively harsh or lenient. Credit discrepancies in multi-mark items may occur for a number of reasons (see table 3), including mismatch between question and mark scheme and poor differentiation between mark points such that a higher score does not represent evidence of greater achievement. For single mark items, credit discrepancies may be the result of setting the cut-off between creditworthy and non-creditworthy responses in an inappropriate place.

References

Crisp, V., & Sweiry, E. (2006). Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research, 28 (2),* 139-154.

Crisp, V., Sweiry, E., Ahmed, A., & Pollitt, A. (2008). Tales of the expected: The influence of students' expectations on question validity and implications for writing exam questions. *Educational Research, 50 (1),* 95-115.

Messick, S. (1989). Validity. In R. Linn (Ed.), Educational measurement, (3rd ed.). Washington, D.C.: American Council on Education.

Pollitt, A. (2009). Abolishing marksism and rescuing validity. Paper presented at the Annual Conference of the International Association for Educational Assessment, Brisbane.

Pollitt, A. and Ahmed, A. (2000). Comprehension Failures in Educational Assessment. Paper presented at the European Conference on Educational Research, Edinburgh.

Pollitt, A, Ahmed, A, Baird, J-A, Tognolini, J, & Davidson, M. (2008). Improving the quality of GCSE Assessment.
Downloaded on 20/08/13 from: http://www2.ofqual.gov.uk/downloads/category/106-gq-monitoring?download=352%3Aimproving-the-quality-of-gcse-assessment-january-2008

Sweiry, E. (2012). Conceptualising and minimising marking demand in selected and constructed response test questions. Paper presented at the Association for Educational Assessment Europe annual conference, Berlin.