This is a single article from *Research Matters: A Cambridge Assessment publication*. http://www.cambridgeassessment.org.uk/research-matters/
© UCLES 2016

Why do so few candidates score 4 out of 8 on this question? The issue of under-used marks in levels-based mark schemes

Sarah Hughes and **Stuart Shaw** Cambridge International Examinations

Introduction

Marks on a question which are rarely achieved by students are 'dead marks' or 'under-used' marks. Under-used marks may have a detrimental effect on reliability and can reduce the discriminative powers of a test (Bramley, 2001). It is necessary to ensure, therefore, that the full range of marks is used.

This study aimed to identify any under-used marks that occur in a History examination for 16 year olds. It explains this occurrence and presents recommendations to ensure that under-used marks are minimalised.

Context

The focus of the study was the Cambridge IGCSE®[1] (International General Certificate of Secondary Education) History Paper 1 (June 2013). The Cambridge IGCSE History syllabus looks at some of the major international issues of the nineteenth and twentieth centuries, as well as covering the history of particular regions in more depth. The emphasis is on both historical knowledge and on the skills required for historical research. Paper 1 contains 25 optional questions. Students are expected to answer three (two questions from Section A – 'Core Content' and one question from Section B – 'Depth Studies'). Each question comprises three parts: a, b and c, with maximum marks of 5, 7 and 8 respectively. The questions in Paper 1 are differentiated by outcome. Student responses are marked using a levels-based mark scheme. The levels of performance in the mark scheme relate to a progression of skills which are summarised in Table 1.

Three issues within the literature are relevant to the under-use of marks in a levels-based mark scheme:

1. The number and width of levels

The number of levels and the width of each level in a levels-based mark scheme can enable or hinder accurate ratings (Shaw & Weir, 2007). Shaw and Weir suggest that each mark point must be defined to clearly and unequivocally embody differing and distinct levels. If this is possible, then the more levels there are, the more precise the rating scale will be. However, markers must be able to clearly and consistently distinguish all of the different levels defined. Pollitt (1991) has argued that it is optimistic to even claim five reliable bands of performance (although this will depend on the target ability of the candidature and the construct being assessed).

Table 1: Skills assessed using the levels-based mark scheme

Question part	*Level and marks available*	*Skills rewarded at each level*
Part a	Level 0 (0 marks)	Answer lacking specific contextual knowledge
	Level 1 (1 mark)	Description
	Level 2 (2–5 marks)	
Part b	Level 0 (0 marks)	Answer lacking specific contextual knowledge
	Level 1 (1 mark)	Description/identification
	Level 2 (2–3 marks)	Explanation
	Level 3 (4–7 marks)	
Part c	Level 0 (0 marks)	
	Level 1 (1 mark)	Answer lacking specific contextual knowledge
	Level 2 (2 mark)	Description/identification
	Level 3 (3–5 marks)	Explanation of one side of the argument
	Level 4 (5–7 marks)	Explanation of both sides of the argument
	Level 5 (8 marks)	Evaluation

Ahmed and Pollitt (2011) argue that it is more problematic to distinguish between marks within a level than between levels. Shaw and Weir (2007) report that markers seem to be able to effectively distinguish between three levels of performance within a band. Fowles (2009) found that, where there were many marks in a GCSE English mark scheme, markers under-used the extreme marks in a band, and differences between markers were exaggerated. Fowles concluded that fewer marks in a greater number of levels may result in greater marking consistency.

The levels-based mark schemes in the Cambridge IGCSE History examination paper have between three and six levels, with each level containing up to four marks.

2. Range of performance within a level

Ahmed and Pollitt (2011) argue that decisions about whether a response is very good or very poor (i.e., which level to apply) are easy judgements to make; it is decisions about which mark within a level to apply that are more difficult. Consequently, they propose that a mark scheme should help markers to score consistently those responses that are close to the extremes of a level. This suggests that descriptions at the extremes of the bands would be most useful.

Some levels in the Cambridge IGCSE History mark schemes function as points-based mark schemes. For example, in Level 2 in Part a and Part b questions which award description of events, one mark is awarded for

[1] Cambridge International Examinations offers the International General Certificate of Secondary Education (IGCSE), which is a two-year qualification aimed at 14 to 16-year-olds. The Cambridge IGCSE encourages learner-centred and inquiry-based approaches to learning. It has been designed to develop learners' skills in creative thinking, inquiry and problem-solving, giving learners a sound preparatory basis for the next stage in their education. More than 70 subjects are available for study, and schools may offer any combination of these subjects. In some Cambridge IGCSE subjects, there are two course levels, known as the 'Core Curriculum' and the 'Extended Curriculum'. The 'Extended Curriculum' includes the material from the 'Core Curriculum', as well as additional, more advanced material.

each point given. In these cases the differentiation between each mark within the level is precisely described. In some levels in Part b and Part c questions there is less prescription and markers are required to make a judgement between marks within the level by following the marking guidance: "Where a band of marks is indicated for a level these marks should be used with reference to the development of the answer within that level."

3. A priori versus empirically-derived levels

Levels-based mark schemes for many general qualifications have been developed using an a priori approach based on the judgement and experience of expert syllabus developers and question writers (Lumley, 2002). Alternatively, some mark scale developers propose an empirical approach to developing level descriptors informed by the analysis of actual student performance (e.g., Milanovic & Saville, 1996; Weir, 2003). Upshur and Turner (1995) argue that an empirical method almost certainly guarantees that the whole range of the rating scale is employed thereby eliminating any under-used marks.

Research questions

The two research questions addressed by this study were:

1. Are any marks within the Cambridge IGCSE History paper under-used?
2. What factors impact on the occurrence of under-used marks?

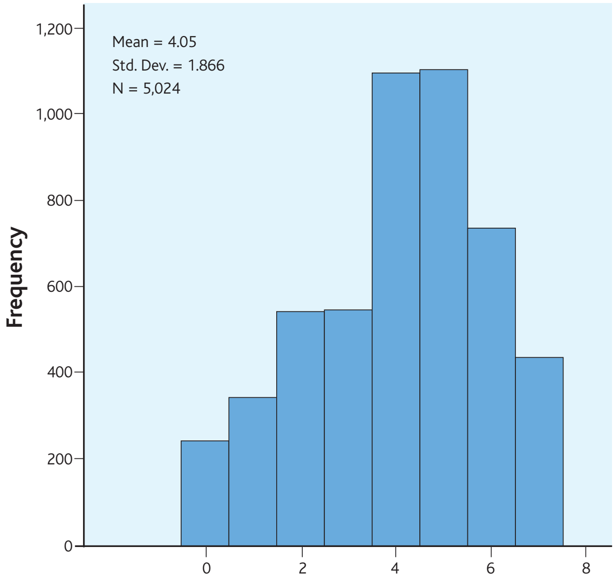

Figure 1: Mark distribution in which no marks are under-used

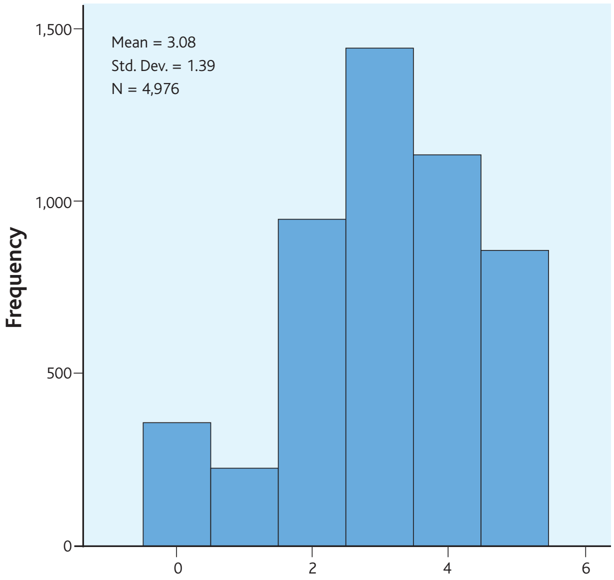

Figure 2: Mark distribution for Part a questions in which a mark of '1' is under-used

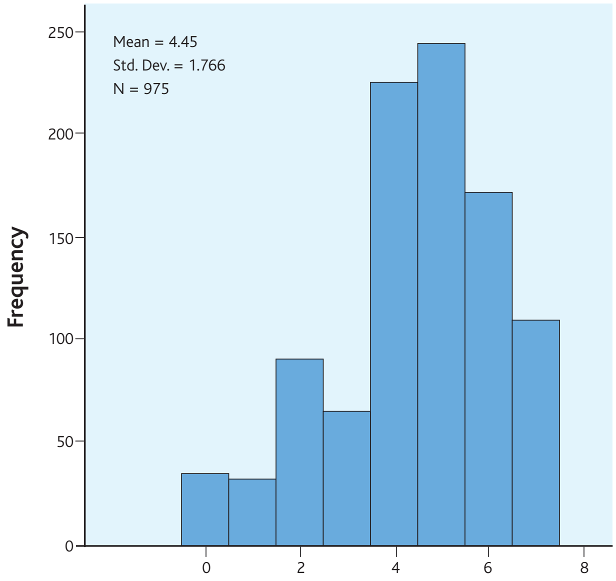

Figure 3: Mark distribution for Part b questions in which marks of '1' and '3' are under-used

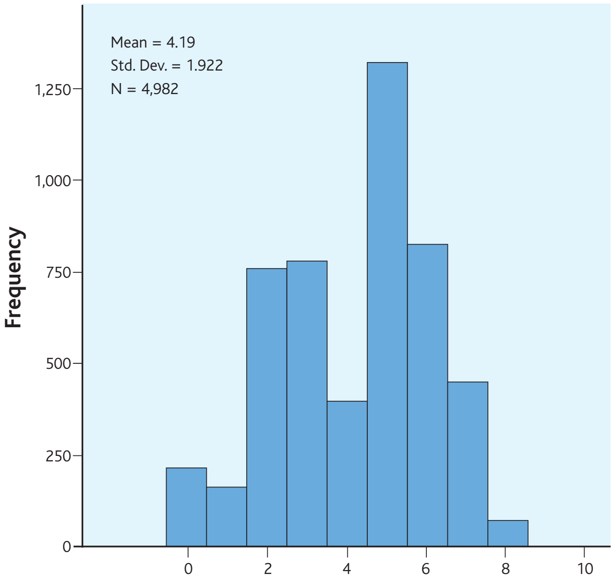

Figure 4: Mark distribution for Part c questions in which a marks of '1' and '4' are under-used

Methodology

Research question 1: Are any marks within the Cambridge IGCSE History paper under-used?

Traditional item analyses

Traditional item analyses were carried out. Analysis included estimates of Backhouse P[2] (a measure of internal consistency using average correlation of items) and histograms showing mark frequency distributions for each of the questions. The data set included 8,144 candidates who took the Cambridge IGCSE History Paper 1 in June 2013.

Rasch analyses

The data was analysed using the Rasch partial credit model (Masters, 1982) with FACETS software (Linacre, 2005). Three separate models were fitted: one examining Part a questions, one looking at Part b and one looking at Part c. Within each of these models, data from any of the 25 optional questions that were answered by at least 100 candidates were included. It is not necessary for every person to have attempted every question for the software to be able to estimate the person and item parameters, but it does require sufficient overlap of persons and questions such that there are no subsets of questions that have only been attempted by a subset of the persons. Separate overall difficulty parameters were estimated for each question. However, across the different questions, the sizes of the differences in difficulties between each successive mark (i.e., the category thresholds) were assumed to be constant. As illustrated later, these category threshold estimates were used to identify potentially under-used marks.

Research question 2: What factors impact on the occurrence of under-used marks?

Repertory Grid analyses

Structured interviews with four interviewees were carried out using the Repertory Grid Technique (Fransella, Bell & Bannister, 2004). This technique identifies the ways that a person construes (interprets or gives meaning to) his or her experience. The Repertory Grid Technique is underpinned by Personal Construct Theory, developed by George Kelly (1955/1991).

Four markers were interviewed either face-to-face or by telephone. Markers were given copies of six examination questions containing under-used marks and asked to consider two exam questions at a time. In order to elicit marker's constructs relating to a number of examination questions with under-used marks, markers were provided with the following prompts (Landfield, 1971):

- *Think of these two exam questions and why the under-used marks were rarely awarded.*
- *Are the two questions alike in terms of why the under-used mark is rarely awarded? If so, how are they alike?*
- *Are the two questions different in terms of why the under-used mark was rarely awarded? If so, how are they different?*

Inductive coding (using codes generated by the researcher) was adopted. Jankowicz (2004) suggests that inductive coding requires that the researcher:

- Identifies themes in the data
- Allocates each segment of data to a theme (or to more than one theme)
- Defines the themes
- Finds examples of each theme
- Finds the frequency of each theme.

Analysis of qualitative data was facilitated using MAXQDA, software for qualitative and mixed methods data analysis.

Findings

Research question 1: Are any marks within the Cambridge IGCSE History Paper under-used?

Traditional item analyses

The measure of internal consistency (Backhouse P) of 0.92 suggests that the questions on the paper are measuring the same construct. Figure 1 shows, for illustration, a mark distribution where no marks are under-used. Figures 2, 3 and 4 show examples of Part a, b and c questions (respectively) which exhibit under-used marks.

These findings were triangulated with those from the Rasch analyses.

Rasch analyses

Not unsurprisingly, score frequencies for the items included in the Rasch analyses (Table 2) show the same pattern or under-use as the mark distributions, that is: a mark of '1' in Part a questions, a mark of '3' in Part b questions, a mark of '4' in Part c questions and possibly a mark of '1' in all question parts.

Table 2: Score frequencies for Part a, b and c questions

Score	*Part a*		*Part b*		*Part c*	
	No.	*%*	*No.*	*%*	*No.*	*%*
0	1,409	6	786	3	816	3
1	977	4	861	4	1,048	4
2	3,188	15	2,149	9	3,576	15
3	5,051	23	2,529	11	3,759	15
4	5,138	23	5,055	21	2,287	9
5	6,195	28	5,384	23	5,599	23
6	-	-	4,526	19	3,782	16
7	-	-	2,529	11	2,983	12
8	-	-	-	-	479	2

'Category Probability Curves' showing the relation between the probability of a given category as a function of person location (in logits) were generated using the FACETS software and are shown in Figures 5, 6 and 7. Figure 5 indicates that for the least able students (with an ability measure on the x axis of -3 logits) the most likely outcome (with a probability of about 0.9) is a mark of '0'. As ability increases, the probability of getting no marks reduces. For students with ability of between about 0 and +0.8 logits, the most probable outcome is a mark of '3'.

Under-used marks are defined here as those which are not the most probable outcome at any point on the ability scale. Adams, Wu and Wilson (2014) propose that marks which are not most probable are not

[2] Backhouse P is a measure of reliability (internal consistency) for tests with optional questions. Values range from 0 to 1, where higher values indicate more reliable tests.

necessarily evidence of a problem, but may be an indication of the relative number of respondents in each category. Nonetheless, Adams et al. (2014) recognise that these may be an indication that an item is not functioning as intended and may indicate issues with the discrimination of the question.

In Figure 5 a mark of '1' is not the most probable outcome for any ability. This is evidence that a mark of '1' is under-used. Figures 5, 6 and 7 show that:

- A mark of '1' is under-used in all question parts (a, b and c) indicating that it is rare for a student to be awarded the one available mark for an 'answer lacking specific contextual knowledge'.
- A mark of '3' is under-used in question Part b. This mark is awarded for description/identification.
- A mark of '4' is under-used in question Part c. This mark is rewarded for an explanation of one side of the argument.

Research question 2: What factors impact on the occurrence of under-used marks?

Four themes were identified by markers as prominent within the data: 1) the skills assessed; 2) marking issues; 3) questions features; and 4) topic content. Frequencies of each theme manifest in the data are shown in Table 3. It is interesting to note from Table 3 that, in terms of references to themes, 'Skills assessed' and 'Question features' were mentioned far more often than 'Marking issues' or 'Topic content'.

Table 3: Frequency of markers' references to themes

Theme	*No.*
1. Skills assessed	
Evaluation	5
Explanation	20
Description/identification	14
Balance of argument	14
2. Marking issues	
Overlapping marks	2
3. Question features	
Question language	12
Familiarity of question type	7
4. Topic content	
Familiarity of topic	2
Question parts	3

1. Skills assessed

A mark of '1' is rewarded for a response 'lacking in specific contextual knowledge'. The reasons for under-use of this mark appear to relate (in part) to marker expectations. Expectations are partly set by knowledge of which question part is being attempted. Part a questions, for example, are described by one marker as containing "less difficult content" and by another as "easier than b or c". Part c questions were described as neither easy nor simple, but as demanding. The following comments were illustrative of this point:

- this is a more difficult area to study,
- a 'sophisticated question' which 'ramps up' from identification skills to explanation skills.

```
  -3.0       -2.0       -1.0        0.0        1.0        2.0        3.0
   ++----------+----------+----------+----------+----------+----------++
 1 |                                                                   |
   |                                                                   |
   |000                                                                |
   |   0000                                                            |
   |       00                                                       555|
 P |         00                                                   55   |
 r |           00                                              555     |
 o |             00                                          55        |
 b |               0                                        5          |
 a |                0                                     55           |
 b |                 00                                 55             |
 i |                   0                               5               |
 l |                    0            33333333        55                |
 i |                     0 2222222**3        3**444**444444            |
 t |                    22*     33  222    444  335        444         |
 y |                  22   00 33       2244     5533          4444     |
   |             111**1111113*        444222  55    333           4444 |
   |       111111222     33311**1  444      **         333            4|
   | 111111   222     333       ***11    555  2222        3333         |
   |1  2222222   33333     44444   ******11       222222      3333333  |
 0 |***********************55555555    0000****************************|
   ++----------+----------+----------+----------+----------+----------++
  -3.0       -2.0       -1.0        0.0        1.0        2.0        3.0
                                  Ability
```

Figure 5: Category Probability Curves for Part a questions

```
   -4.0         -2.0          0.0          2.0          4.0          6.0
   ++------------+------------+------------+------------+------------++
 1 |                                                                  |
   |                                                                 7|
   |                                                             7777 |
   |00                                                         77     |
   |  00                                                     77       |
 P |    0                                                  77         |
 r |     0                                                7           |
 o |      0                                             77            |
 b |       0                                           7              |
 a |        0                                         7               |
 b |         0                                66666  7                |
 i |          0              4444    5555   66     6*                 |
 l |               222     44    4455    5*6       7 66               |
 i |           0 22   22  4      5544    6 55     7    66             |
 t |         111*      3**33    5    4  6    5  77       6            |
 y |      111 22 *1  33 42  33 5      *6      5*          66          |
   |    11   2    01*  4  22 5*3     6 44     7 5           66        |
   | 111   22     3* 1*     *   3  66    44 77   55           666     |
   |1    22     33  ** 11 55 22  **3      7*4      555           6666 |
   | 2222    333 444  0***11  6**2  33**77   4444     5555           6|
 0 |******************************************************************|
   ++------------+------------+------------+------------+------------++
  -4.0         -2.0          0.0          2.0          4.0          6.0
                                  Ability
```

Figure 6: Category Probability Curves for Part b questions

```
  -6.0     -4.0     -2.0      0.0      2.0      4.0      6.0      8.0
   ++--------+--------+--------+--------+--------+--------+--------++
 1 |                                                                |
   |00                                                             8|
   |  00                                                        888 |
   |    00                                                    88    |
   |      0                                                  8      |
 P |       0                                                8       |
 r |        0                                  7777        8        |
 o |         0                               77    7      8         |
 b |                                        7       7    8          |
 a |          0                            7         77 8           |
 b |                222         555       7            *            |
 i |           0   2   22      5   5   6 7            8             |
 l |            0 2      333  5     566 66           8  7           |
 i |             2      32  3      665  7 66        8    7          |
 t |           1**     3  2  *    6   57    6      8      77        |
 y |         11   *1  3    25 3  6    75     6    8         7       |
   |       11  2    1*      44446    7  5     6  8           7      |
   |     11   2    03 1   4*2  *44  7    5     6*             77    |
   |  111   22    3300 1*45  26 334*      55  88 66             777 |
   |11   222    33   4**5*1 662277*3444   88**5    6666            7|
 0 |****************************************************************|
   ++--------+--------+--------+--------+--------+--------+--------++
  -6.0     -4.0     -2.0      0.0      2.0      4.0      6.0      8.0
                                Ability
```

Figure 7: Category Probability Curves for Part c questions

Very few candidates provided an answer 'lacking specific contextual knowledge' and markers indicated that in Part a questions they expect candidates to achieve the highest available level (Level 2). Such an expectation may contribute to the under-use of a mark of '1' in Level 1.

The under-use of a mark of '3' in Part b questions is described as relating to progression from mark scheme Level 2 – which rewards students' ability to describe or identify, to Level 3 – which rewards students' ability to explain.

ABILITY TO DESCRIBE OR IDENTIFY

The stimulus material for one question is a picture of Hungarian refugees fleeing from Austria after the failure of an uprising. Markers reported that candidates find it easy to identify with people involved in historical events leading them to describe participants' experience of those events, rather than explain or evaluate the impact of those events. However, even students working at Level 2 (identifying and/or describing) are highly likely to achieve one mark in Level 3. This is because, when providing a narrative, students typically gain one explanation mark raising their performance to the bottom of Level 3. Markers will usually manage to identify some explanation in a response that is mainly descriptive. For example, one marker described how candidates might provide significant amounts of 'floundering description' and 'happen across' one explanation mark moving them from Level 1. This indicates that markers can recognise explanation in what is mainly a descriptive answer. Markers also described how this may account for students achieving a mark of '4' in Level 3 (in Part b questions) leaving a mark of '3' in Level 2 under-used.

ABILITY TO EXPLAIN

Markers described two features of questions which help candidates move from providing a descriptive response to providing an explanation:

1. Some topic areas foster explanatory responses. Markers described some topic areas which lend themselves to explanation, for example, The Cold War or the advantages of Stalin's economic policies. Such topics prompt explanatory rather than descriptive responses from candidates.
2. The teaching of higher order skills in some schools and colleges with particular focus on explanation and evaluation facilitates progression through the levels. One marker explained how some schools and colleges support candidates' progression through levels by teaching "the difference between description and explanation". Understanding the difference between description and explanation on the part of the candidate supports progression to Level 3 in Part b questions (for a mark of '4') and could, therefore, reduce the occurrence of a mark of '3' in Level 2.

ABILITY TO PROVIDE A BALANCED ARGUMENT

The Part c mark scheme contains five levels. In particular, Level 3 ('3'–'5' marks) rewards explanation of one side of the argument; Level 4 ('5'–'7' marks) rewards explanation of both sides of the argument. Whether a response includes a balanced argument or not appears to be the most significant factor in the under-use of a mark of '4' in Part c questions. Whilst not all students necessarily provide a balanced or two-sided argument, they are able to refer to both sides of the argument (even if one side is weak) and so achieve marks in Level 4, leaving at least one mark in Level 3 under-used. Markers explained why even an unbalanced argument is unlikely to be totally-one sided:

- *Teaching of both sides of an argument:* Students are likely to be trained to address both sides of any argument (but despite training, in an exam situation students may forget to focus on both sides). Each side of an argument may be given different treatment by teachers. For example, one marker said that "they teach Germany better than the Soviet Union". This may be because there are more knowledge or resources available to teach one context compared to another, or because contexts may differ in their complexity.
- *Distinct sides of an argument:* Where the information relating to the two different sides of the argument is distinct, markers reported that candidates are better able to provide a two-sided argument. One marker, for example, pointed to the clear benefits of the Nazi-Soviet pact to Germany on the one hand and the Soviet Union on the other.
- *Stimulus material:* Most stimulus materials are balanced in the view they present enabling candidates to see the two sides of the argument and achieve marks in Level 4 (giving a two-sided argument). Occasionally, stimulus material was described as having a bias towards one side of the argument (e.g., a quote from Hitler rallying his generals in May 1939 or Anthony Eden supporting the League of Nations).
- *Question wording:* Question wording supports candidates' engagement with both sides of the argument. In only one question did markers identify question wording which could be biased towards one side of an argument.

2. Marking issues

OVERLAPPING MARKS

A mark of '5' in Part c questions is an overlapping mark and is gained at the top of Level 3 or at the bottom of Level 4. Markers suspected that a mark of '5' in Level 3 is under-used in the same way as a mark of '4' in Level 3 is under-used . One marker estimated that about 80 per cent of students gaining a mark of '5' are doing so in Level 4. This suggests that, although a mark of '5' was not under-used overall, it may be under-used in Level 3. Of the three marks available ('3', '4' or '5') in Level 3, it is likely that both a mark of '4' and a mark of '5' are under-used. Since this study, the overlapping mark has been removed from the mark scheme.

The finding that both a mark of '4' and a mark of '5' are under-used[3] may shift the focus from the under-use of a single mark to the under use of Level 3 as a whole. Level 3 rewards candidates who present a developed one-sided argument and the findings suggest that students rarely give a developed one-sided argument.

3. Question features

- *Content compatible/obligatory language: Content-obligatory* language is content- or discipline-specific and academic in nature and it is necessary for learning key concepts in the subject (Fortune & Tedick, 2008). *Content-compatible* language goes beyond the student's subject learning. In the Cambridge IGCSE History paper, for example, the non-historical language of the question stem and instructions is *content-compatible*. *Content-compatible* language is uncomplicated and enables candidates to access the higher levels of the mark scheme and so leaves some marks under-used. Conversely (and much more rarely in the Cambridge IGCSE History paper)

[3] Item Level Data does not differentiate between the two routes to a mark of '5'.

complex language provides a barrier to progression through the levels.

- *Familiarity of question types:* For a topic area that is regularly set, the question writer needs to devise new topic-related questions which are as yet unseen. Three of the four markers described how this may lead to "obscure", "sophisticated" or "unfamiliar" question types. This is problematic because the setter needs to devise novel questions which are neither obscure nor overly sophisticated. Questions which were described as set "in a new way" could hinder a candidate's ability to show his or her skills and thereby impede progress through the levels.

4. Topic content

- *Topic familiarity:* Markers made frequent references to familiar topic content (as opposed to unfamiliar topic content). For example, those which are "regularly" set topics which "occur every year", are "similar to past questions", are "well taught" and candidates "thoroughly know it". Another marker described how candidates "will have had experience of [the topic] because past papers include it and it is probably well prepared for". Findings suggest that familiar topics tend to result in better quality answers than new or rarely assessed content areas, and that familiar content allows candidates to move up the levels in the mark scheme taking them beyond performance that might occur with a less familiar topic.
- *Question parts:* Certain topics are associated with particular question parts. The question type may be unfamiliar to candidates because content usually assessed using a Part a or a Part b question, for example, is assessed using a Part c type question. This changes the skills being assessed in relation to content: Part c questions require candidates to provide a balanced argument and to evaluate the argument, neither of which are required in Part a or b questions.

Conclusion

In some cases the under-use of marks might be prompted by the accessibility of the questions (in the form of familiar topics and question types, predictable skills, simple non-historical language and clear and readable stimulus material). These features enable students to perform at their potential level of ability without being distracted by irrelevant (non-historical) demands in the questions, and so leave marks in lower levels under-used.

Accessible questions are not necessarily 'easy' questions; accessible questions can assess high-level skills and demanding content. What makes a question accessible is that it assesses the target skills and content without assessing factors irrelevant to the intended construct(s) (e.g., question wording or layout). Accessibility is desirable (in that it minimises 'construct irrelevant variance').[4] If a consequence of accessibility is that some marks are under-used, does it matter?

Adams et al. (2014) use the term 'middle score categories' for marks in a question that appear in the middle of the available mark range. Where few respondents achieve middle score categories, these marks are not useful and may indicate issues with the discrimination of the question (Adams et. al., 2014). As such, under-used marks may threaten validity.

This study shows that two of the three marks available in one of the mark scheme levels for the Cambridge IGCSE History Paper 1 are under-used. This raises questions about the purpose of this level and the validity of a mark scheme level which is rarely awarded. An empirical approach to developing level descriptors based on student performance, rather than a declared construct of performance (Upshur & Turner, 1995) could help reduce the number of under-used marks.

The study also suggests a need for clarification of the level descriptors which would support examiners making judgements between single marks (Ahmed & Pollitt, 2011; Shaw & Weir, 2007). Special attention also needs to be given to the setting of questions on any over-exposed topics which require novel questions to prevent repetition over time.

Research relating to levels-based mark scheme development and application would suggest a number of features that characterise effective practice which inform the construction and continuing improvement of general qualifications such as the Cambridge IGCSE History mark scheme. Almost all best practice described in the literature is already applied to the Cambridge IGCSE History mark schemes:

- using positively worded levels (Galaczi, ffrench, Hubbard & Green, 2011);
- providing indicative content (Tisi, Whitehouse, Maughan & Burdett, 2013);
- having a number of levels that markers can effectively distinguish (Ahmed & Pollitt, 2011);
- articulating clear and precise definitions of the distinction between different levels and between marks within a level (Shaw & Weir, 2007);
- reducing the use of relative adjectives (e.g., very frequent, fairly frequent, some) to differentiate descriptions of performance (Galaczi et. al., 2011);
- including examiner training and standardisation as part of the marking process (Baird, Beguin, Black, Pollitt & Stanley, 2011);
- ensuring expectations are made clear to students and teachers about the skills being assessed and the assessment model used to assess them (Sweiry, Crisp, Ahmed & Pollitt, 2002).

As a future line of research inquiry, one potential area of interest relates to the use of empirical evidence to establish the construct being assessed in each level (Upshur & Turner, 1995). This practice is not generally employed in the development of mark schemes for general qualifications as there is no pre-testing of the papers. However, it may be possible in an examination like Cambridge IGCSE History which uses similar mark scheme structures over time, to analyse student performance in one year and apply lessons to future papers and mark schemes.

The research reported here highlights concerns which were already articulated by senior examiners and which resulted in a re-designed mark scheme for the Cambridge IGCSE History paper ready for the June 2015 examination. The new mark scheme aims to support examiners judging the quailty of answers at the top of Levels 3 and 4 in Part c questions. The revised mark scheme has eliminated overlapping marks. Further research could usefully focus on monitoring student outcomes in Levels 3 and 4 in the future to evaluate this new mark scheme structure.

[4] Variability in performance which is not attributable to the construct being assessed.

References

Adams, R. J., Wu, M. L., and Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement*, *72*(4), 547–573.

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

Baird, J. A., Beguin, A., Black, P., Pollitt, A. and Stanley, G. (2011). The Reliability Programme: Final Report of the Technical Advisory Group. In: *Ofqual's Reliability Compendium* (Chapter 20). Coventry: The Office of Qualifications and Examinations Regulation.

Bramley, T. (2001) The Question Tariff Problem in GCSE Mathematics. *Evaluation and Research in Education*, *15*(2), 95–107.

Fortune, T. W. and Tedick, D. J. (Eds.) (2008) *Pathways to multilingualism: Evolving perspectives on immersion education*. Clevedon, England: Multilingual Matters, Ltd.

Fowles, D. (2009) 'How reliable is marking in GCSE English?' *English in Education*, *43*(1), 49–67.

Fransella, F. Bell, R. and Bannister, D. (2004) *A Manual for Repertory Grid Technique (2nd Edition)* Chichester, UK: John Wiley and Sons.

Galaczi, E. D., ffrench, A., Hubbard, C. and Green, A. (2011) Developing assessment scales for large-scale speaking tests: a multiple-method approach, *Assessment in Education: Principles, Policy and Practice*, *18*(3), 217–237.

Jankowicz, D. (2004) *The Easy Guide to Repertory Grids*. Chichester, UK: John Wiley and Sons.

Kelly, G. A. (1955/1991) *The Psychology of Personal Constructs*. London, UK: Routledge.

Landfield, A. W. (1971) *Personal Construct Systems in Psychotherapy*. Chicago, USA: Rand McNally.

Linacre, J. M. (2005). *A User's Guide to FACETS Rasch-Model Computer Programs*. Available at: www.winsteps.com

Lumley, T. (2002) Assessment criteria in a large-scale writing test: What do they really mean to raters? *Language Testing*, *19*(3), 246–76.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika* *20*(2), 149–174.

Milanovic, M. and Saville, N. (1996) Performance testing, cognition and assessment: selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem. *Studies in Language Testing 3*. Cambridge, UK: University of Cambridge Local Examinations Syndicate/Cambridge University Press.

Pollitt, A. (1991). Response to Alderson, Bands and scores. In J. C. Alderson & B. North (Eds.), *Language Testing in the 1990s*. London, UK: MacMillan.

Shaw, S. D. and Weir, C. J. (2007) Examining Second Language Writing: research and practice. *Studies in Language Testing 26*. Cambridge, UK: University of Cambridge Local Examinations Syndicate/Cambridge University Press.

Sweiry, Z., Crisp, V., Ahmed, A, and Pollitt, A. (2002) *Tales of the Expected: The Influence of Students' Expectations on Exam Validity*. British Educational Research Association Conference, Exeter, UK, September 2002.

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

Upshur, J. and Turner, C. (1995) Constructing rating scales for second language tests. *ELT Journal*, *49*(1), 3–12.

Weir, C. J. (2003) A survey of the history of the Certificate of Proficiency in English (CPE) in the twentieth century. In C.J. Weir & M. Milanovic (Eds.). Continuity and Innovation: A History of the CPE Examination 1913–2002. *Studies in Language Testing 15*. Cambridge, UK: University of Cambridge Local Examinations Syndicate/Cambridge University Press.

Maintaining test standards by expert judgement of item difficulty

Tom Bramley Research Division and **Frances Wilson** OCR (The study was completed when the second author was based in the Research Division)

Introduction

This article describes two methods for using expert judgements about examination questions (items) to arrive at a cut-score (grade boundary) on a new examination paper where none of the items has been pre-tested. We wanted to see if we could exploit the wealth of data about item difficulty that has been available in the years since the majority of papers have been marked (scored) on-screen.

The General Certificate of Secondary Education (GCSE) and the General Certificate of Education GCE Advanced level (A level) are high-stakes curriculum-based examinations taken at age 16 and 18 respectively by pupils in England. They are offered by three Awarding Organisations (AOs), and schools can decide which AO's exams they enter their pupils for. Outcomes are reported on a grade scale (A* to G at GCSE; A* to E at A level, with U indicating 'ungraded' for both). From 2017, reformed GCSEs in England will be graded on a 1–9 scale. The full assessments normally consist of several components (e.g., written examination papers, practical or coursework assessment, portfolios, speaking tests, musical performances etc.). The assessments are usually graded at component level, and the overall grade is determined by aggregation rules which can vary considerably depending on the structure of the assessment (e.g., whether the assessment is 'linear', where all components are taken at the end of the course, or 'modular', where assessment units can be taken at various stages throughout the course). At component level, the grading process involves establishing the cut-scores (grade boundaries) on the raw mark scale that define the ranges of raw scores mapping to each grade.[1] A regulatory code of practice (Office of Qualifications and Examinations Regulation [Ofqual], 2011) sets out the mandatory aspects of this process, which requires the AOs to consider a variety of sources of evidence. Benton and Bramley (2015) show that these sources of evidence can be broadly classified as: i) evidence about the ability of the cohort of examinees; ii) evidence about the difficulty of the examination; and iii) evidence about the quality of work produced in the examination.

Setting the grade boundaries is essentially a standard-maintaining process (as opposed to a standard-setting process) where the aim is for

[1] Only particular 'key boundaries' are established by the 'Awarding Committee' – the other boundaries are derived from these by interpolation rules. At A level, the key boundaries are at grades A and E.