Cambridge Assessment

# The usefulness of detailed marks within the levels of levels-based mark schemes

*Research Report*

Sylwia Macinska & Tom Benton

12 March 2020

## Author contact details:

Sylwia Macinska & Tom Benton
Assessment Research and Development,
Research Division
Cambridge Assessment
The Triangle Building
Shaftesbury Road
Cambridge
CB2 8EA
UK

Macinska.S@cambridgeassessment.org.uk
Benton.T@cambridgeassessment.org.uk

http://www.cambridgeassessment.org.uk

As a department of Cambridge University, Cambridge Assessment is respected and trusted worldwide, managing three world-class examination boards, and maintaining the highest standards in educational assessment and learning. We are a not-for-profit organisation.

## How to cite this publication:

Macinska, S. & Benton, T. 2020. *The usefulness of detailed marks within the levels of levels-based mark schemes*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

## Introduction

Accurate marking is critical to the integrity of any assessment. There are many factors that can influence the quality of marking, one of which is mark scheme design (Child, Munro & Benton, 2015). Mark schemes provide a structure to the marking process, in an attempt to ensure that all the markers follow exactly the same instructions and (to the extent that this is possible) make the same judgements when evaluating candidates' work (Alpha Plus, 2014).

The mark scheme is an integral component of an item and should be developed in tandem with the item itself (Ofqual, 2011). Based on the nature of the item they are describing, mark schemes can be categorised into objective, point-based and levels-based (Massey & Raikes, 2006). Objective mark schemes are typically used for one or two-word responses where an unambiguously correct answer can be identified. Point-based mark schemes list key words, statements or ideas, with a mark awarded for each point candidate makes that matches the response listed in the mark schemes. Levels-based mark schemes are used for items that require an extended written response (e.g. essay). This type of items typically carries a high mark tariff, which is divided into smaller bands called levels. The levels-based mark scheme contains descriptions of the standard of response required to achieve each level.

To help ensure the reliability of the assessment, the mark scheme should be rigorous and consistently applied. However, research shows that the demand of the marking task differs based on the type of the mark scheme adopted in the assessment, which in turn affects marking accuracy (Black, Suto & Bramley, 2011). Levels-based mark schemes typically suffer from lower marking accuracy than point-based or objective mark schemes (Tisi, Whitehouse, Maughan & Burdett, 2013). It has been argued that this is due to a higher cognitive load associated with levels-based marking; extended responses are likely to require more interpretation from the examiners and are more likely to include unexpected content, rendering the marking process more effortful (Bramley, 2009;  Suto & Nadas, 2008). Simple marking strategies such as matching can no longer be used with levels-based mark schemes, requiring more complex approaches such as evaluating and scrutinising (Suto & Greatorex, 2008).

Levels-based marking typically follows a two-stage process (Pinot de Moira, 2011; Hughes, 2018). Firstly, an initial assessment of a candidate response is made, with the examiner classifying the response into a single level within the mark scheme. Secondly, this judgement is further refined and the examiner decides on the mark within the selected level band. As the type of items for which levels-based mark schemes are used allows for a range of acceptable responses, having a detailed points-based mark scheme or a prescriptive set of responses would not be appropriate nor practical (Bramley, 2009; Pinot de Moira, 2011). Consequentially, levels-based marking relies predominantly on examiner's judgement. Clear descriptions of the standard of response required to achieve each level are therefore critical for markers to be able to distinguish between all levels within the mark scheme (Ahmed and Pollitt, 2011).

The number of levels within the mark scheme and the width of each level band can also affect the accuracy of levels-based marking (Shaw & Weir, 2007). Ahmed and Pollitt (2011) argued that it is easier to decide about the overall quality of work to classify it into an appropriate level than it is to make a fine-grained judgement about which mark within a level band should be applied. Evidence shows that examiners tend to under-use extreme marks within a level, often leaving some marks with the mark scheme unused (Hughes & Shaw, 2016; Pinot de Moira, 2011). Fowles (2009) suggested that greater marking consistency could be achieved by having more levels of response but with fewer marks available within each level.

A simple question that has not been answered by previous research, is whether the detailed marks within the levels are actually useful in levels-based marking at all. In other words how much is gained by asking markers to make fine-grained distinctions within levels? Indeed, it could be the case that distinguishing between levels of responses provides a sufficiently accurate picture of candidate performance without any further detail being necessary.

### Aim

The aim of this project was to explore whether using detailed marks within the levels-based mark schemes provided by markers leads to more accurate results than using the levels within the mark scheme alone. More specifically, following Benton and Gallacher (2018), the focus was on the relative predictive value of detailed marks as opposed to simply the level bands within the mark scheme. Predictive value was measured as the level of correlation between each of these measures and external measures of candidates' performance. In theory, detailed marks within the levels should lead to an improved predictive value as they provide more fine-grained judgement about the quality of candidates' responses. However, if markers tend to under-use extreme marks within the levels and, in general, find it more difficult to distinguish between different performances within the levels, they may introduce unnecessary noise into the process. In such cases, the predictive value of the simple levels within the mark scheme should be at least as high as that of the detailed marks.

## Method

### Data

The data was drawn from a selection of OCR essay items that were marked using levels-based mark schemes. Manual search was used to identify the essay items, with the aim to include a variety of subjects with relatively large entries. The essay items came from three GCSE and three A-level qualifications taken from exam series between the years 2013-2016. The list of the subjects used in the current project and the number of marks available on the essays used for analysis in each case is presented in Table 1.

*Table 1 The selection of essay items used in the project.*

| Qualification | Sessions | Maximum mark for question(s) analysed |
|---|---|---|
| A level Subject 1 | June 2013-2015 | 30 |
| A level Subject 2 | June 2013-2015 | 15 |
| A level Subject 3 | June 2013-2015 | 48 |
| GCSE Subject 4 | June 2013-2015 | 26 |
| GCSE Subject 5 | June 2013-2015 | 16 |
| GCSE Subject 6 | June 2015-2016 | 16 |

For each selected item, the item scores were collated from OCR database. The mark schemes were consulted to collapse the item marks back into levels, resulting in two measures of performance for each analysed item: (1) achieved item mark and (2) level of response. There was variation in mark schemes as to whether the highest or lowest level was described first. If needed, the levels within the mark scheme were renumbered so that higher numbers represented a superior performance. The mark schemes had between four and nine levels, with each level containing up to seven marks. In some mark schemes, the mark of zero was included in the lowest level and in others, it was identified separately outside the levels (see Table 2).

*Table 2 The number of levels within the marks scheme and width of level bands for each qualification. 'Zero included' refers to occurrences where the mark of zero was included within the lowest level. 'Zero excluded' refers to occurrences where an additional level was created for the mark of zero.*

| **Qualification** | **Levels within mark scheme** | **Marks within each level band** |
|---|---|---|
| A level Subject 1 | 6 (zero included) | 6-5 |
| A level Subject 2 | 4 (zero excluded) | 4-3 |
| A level Subject 3 | 7 (zero included) | 7 |
| GCSE Subject 4 | 9 (zero included) | 3 |
| GCSE Subject 5 | 4 (zero excluded) | 4 |
| GCSE Subject 6 | 6 (zero excluded) | 3-2 |

### Measures of performance

Three external measures of performance were identified for each item to evaluate the predictive value of marking using levels within the mark scheme versus detailed marks. These included:

1. the mark on the remainder of the exam paper excluding the item being analysed – (remaining marks measure = total mark on the test – item marks)
2. the mark on another exam paper within the same qualification taken by a large number of relevant candidates (another unit measure)
3. the Uniform Marks Scale (UMS) marks from an entirely different OCR qualification taken by a large number of the relevant candidates (another qualification measure)

Suitable external units and qualifications were identified via a manual search. For the majority of qualifications, a mixture of the above measures was used (the exception was GCSE Subject 4 where there was no other examined unit within the qualification).

## Results

The Spearman rank order correlations of both detailed marks and levels from the item being analysed were calculated with each of the external measures of performance for each qualification. These are shown in Tables 3 to 8. The same results are also visualised in Figure 1 (items from A level qualifications) and Figure 2 (items from GCSE qualifications).

The item marks within the levels were found to have higher positive correlations with other measures of performance than item levels. Although the magnitude of the difference was very small (median difference between correlations of just 0.03), this pattern of results was observed for all qualifications explored, regardless of the external measure of performance used. The only exception was a correlation between A level Subject 3 marks with another qualification measure for June 2014 series, where the associations between the marks within the levels and levels within the mark scheme did not differ. However, the sample size for this particular analysis was relatively low (N=161).

It is important to note that the external measures of performance significantly correlated with both detailed marks within the levels and the levels themselves. However, the correlations obtained for marks within the levels were stronger. The test of the equality between two dependent correlations was used to evaluate whether these differences were statistically significant (Lee & Preacher, 2013). The results (not shown) confirmed that for the majority of qualifications across all series examined (41 out of the 45 pairs of correlations analysed) the correlations of marks with external measures were significantly higher than correlations of levels with external measures. The exceptions were likely observed due to low samples sizes.

The highest correlations between both marks and levels were typically found with remaining marks on the test measure, followed by another unit measure and then another qualification measure. The results were fairly consistent across the series. This is unsurprising as the remainder of the same test is likely to contain the most similar content and, in addition, was examined on the same day. It may also reflect the so-called “halo effect” where examiners that are positive about a candidates response to one element of an exam may be more generous in marking later items (Ofqual, 2014). This possibility is one reason why multiple external measures of performance were used in this research.

Higher correlations were typically found for qualifications with a higher number of levels within the mark scheme. For example, the correlations between the detailed marks within the mark scheme and external measures were in the range of 0.52 and 0.61 for GCSE Subject 4, which has nine levels within the mark scheme but in the range of 0.32 and 0.48 for A level Subject 2, which has five levels within the mark scheme. Similar patterns were observed for other qualifications. This may simply reflect that items with smaller numbers of levels tended to relate to shorter tasks and also have fewer available marks.

***Table 3 The Spearman rank order correlations of item marks and level within the mark scheme with external measures of performance for A level Subject 1.***

| A level Subject 1 | | | | | | |
|---|---|---|---|---|---|---|
| Measure<br>Session | Remaining Marks | | Another Unit | | Another Qualification | |
| | Marks | Levels | Marks | Levels | Marks | Levels |
| June 2013 | 0.88<br>N = 10260 | 0.83<br>N = 10260 | 0.51<br>N = 593 | 0.48<br>N = 593 | 0.58<br>N = 960 | 0.50<br>N = 960 |
| June 2014 | 0.89<br>N = 10503 | 0.85<br>N = 10503 | 0.48<br>N = 1379 | 0.45<br>N = 1379 | 0.59<br>N = 930 | 0.54<br>N = 930 |
| June 2015 | 0.88<br>N = 11468 | 0.84<br>N = 11468 | 0.52<br>N = 1436 | 0.50<br>N = 1436 | 0.59<br>N = 998 | 0.56<br>N = 998 |

***Table 4 The Spearman rank order correlations of item marks and level within the mark scheme with external measures of performance for A level Subject 2.***

| A level Subject 2 | | | | | | |
|---|---|---|---|---|---|---|
| Measure<br>Session | Remaining Marks | | Another Unit | | Another Qualification | |
| | Marks | Levels | Marks | Levels | Marks | Levels |
| June 2013 | 0.48<br>N = 5530 | 0.42<br>N = 5530 | 0.39<br>N = 3603 | 0.34<br>N = 3603 | 0.32<br>N = 230 | 0.22<br>N = 230 |
| June 2014 | 0.39<br>N = 7200 | 0.33<br>N = 7200 | 0.33<br>N = 7120 | 0.28<br>N = 7120 | 0.37<br>N = 281 | 0.35<br>N = 281 |
| June 2015 | 0.33<br>N = 5634 | 0.28<br>N = 5634 | 0.32<br>N = 5634 | 0.28<br>N = 5634 | 0.41<br>N = 204 | 0.36<br>N = 204 |

***Table 5 The Spearman rank order correlations of item marks and level within the mark scheme with external measures of performance for A level Subject 3.***

| A level Subject 3 | | | | | | |
|---|---|---|---|---|---|---|
| Measure<br>Session | Remaining Marks | | Another Unit | | Another Qualification | |
| | Marks | Levels | Marks | Levels | Marks | Levels |
| June 2014 | 0.68<br>N = 8375 | 0.64<br>N = 8375 | 0.40<br>N = 7272 | 0.38<br>N = 7272 | 0.23<br>N = 161 | 0.23<br>N = 161 |
| June 2015 | 0.65<br>N = 7873 | 0.61<br>N = 7873 | 0.44<br>N = 6767 | 0.41<br>N = 6767 | 0.42<br>N = 170 | 0.36<br>N = 170 |

*Table 6 The Spearman rank order correlations of item marks and level within the mark scheme with external measures of performance for GCSE Subject 4.*

| GCSE Subject 4 | | | | |
|---|---|---|---|---|
| Measure<br>Session | Remaining Marks | | Another Qualification | |
| | Marks | Levels | Marks | Levels |
| June 2013 | 0.61<br>N = 16314 | 0.60<br>N = 16314 | 0.60<br>N = 12909 | 0.58<br>N = 12909 |
| June 2014 | 0.59<br>N = 14636 | 0.57<br>N = 14636 | 0.57<br>N = 11434 | 0.56<br>N = 11434 |
| June 2015 | 0.56<br>N = 13377 | 0.54<br>N = 13377 | 0.52<br>N = 10315 | 0.51<br>N = 10315 |

*Table 7 The Spearman rank order correlations of item marks and level within the mark scheme with external measures of performance for GCSE Subject 5.*

| GCSE Subject 5 | | | | | | |
|---|---|---|---|---|---|---|
| Measure<br>Session | Remaining Marks | | Another Unit | | Another Qualification | |
| | Marks | Levels | Marks | Levels | Marks | Levels |
| June 2013 | 0.45<br>N = 5044 | 0.43<br>N = 5044 | 0.49<br>N = 4976 | 0.45<br>N = 4976 | 0.48<br>N = 356 | 0.46<br>N = 356 |
| June 2014 | 0.43<br>N = 22624 | 0.41<br>N = 22624 | 0.41<br>N = 22587 | 0.39<br>N = 22587 | 0.30<br>N = 1016 | 0.28<br>N = 1016 |
| June 2015 | 0.55<br>N = 22922 | 0.52<br>N = 22922 | 0.51<br>N = 22695 | 0.47<br>N = 22695 | 0.35<br>N = 1136 | 0.31<br>N = 1136 |

*Table 8 The Spearman rank order correlations of item marks and level within the mark scheme with external measures of performance for GCSE Subject 6.*

| GCSE Subject 6 | | | | | | |
|---|---|---|---|---|---|---|
| Measure<br>Session | Remaining Marks | | Another Unit | | Another Qualification | |
| | Marks | Levels | Marks | Levels | Marks | Levels |
| June 2015 | 0.65<br>N = 27491 | 0.62<br>N = 27491 | 0.66<br>N = 13379 | 0.64<br>N = 13379 | 0.61<br>N = 1257 | 0.59<br>N = 1257 |
| June 2016 | 0.64<br>N = 28924 | 0.61<br>N = 28924 | 0.65<br>N = 14399 | 0.62<br>N = 14399 | 0.56<br>N = 1234 | 0.53<br>N = 1234 |

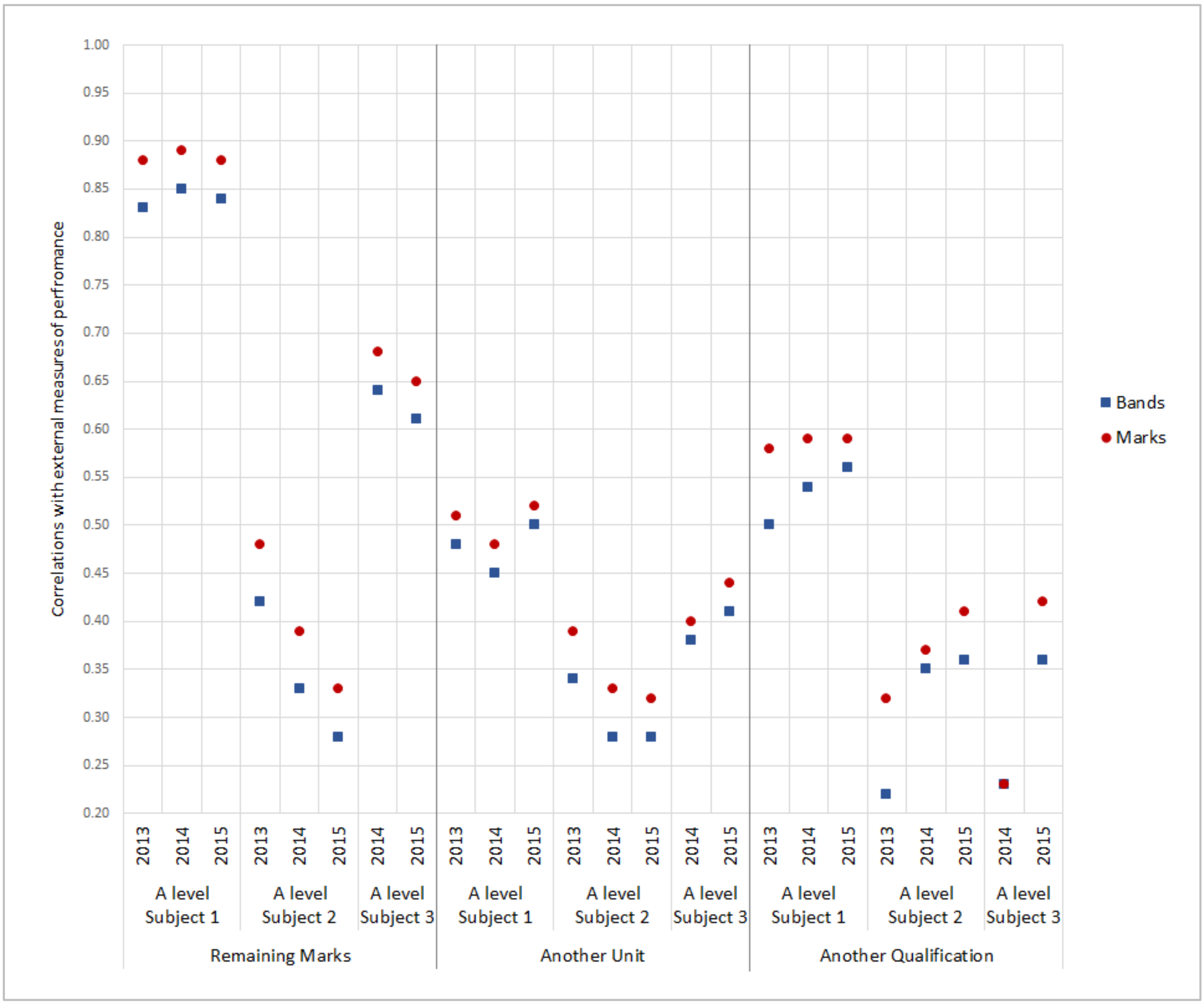

*Figure 1 The Spearman order rank correlation values for levels within the mark scheme (bands) and marks within the bands (marks) with external measures of performance for A level items explored.*

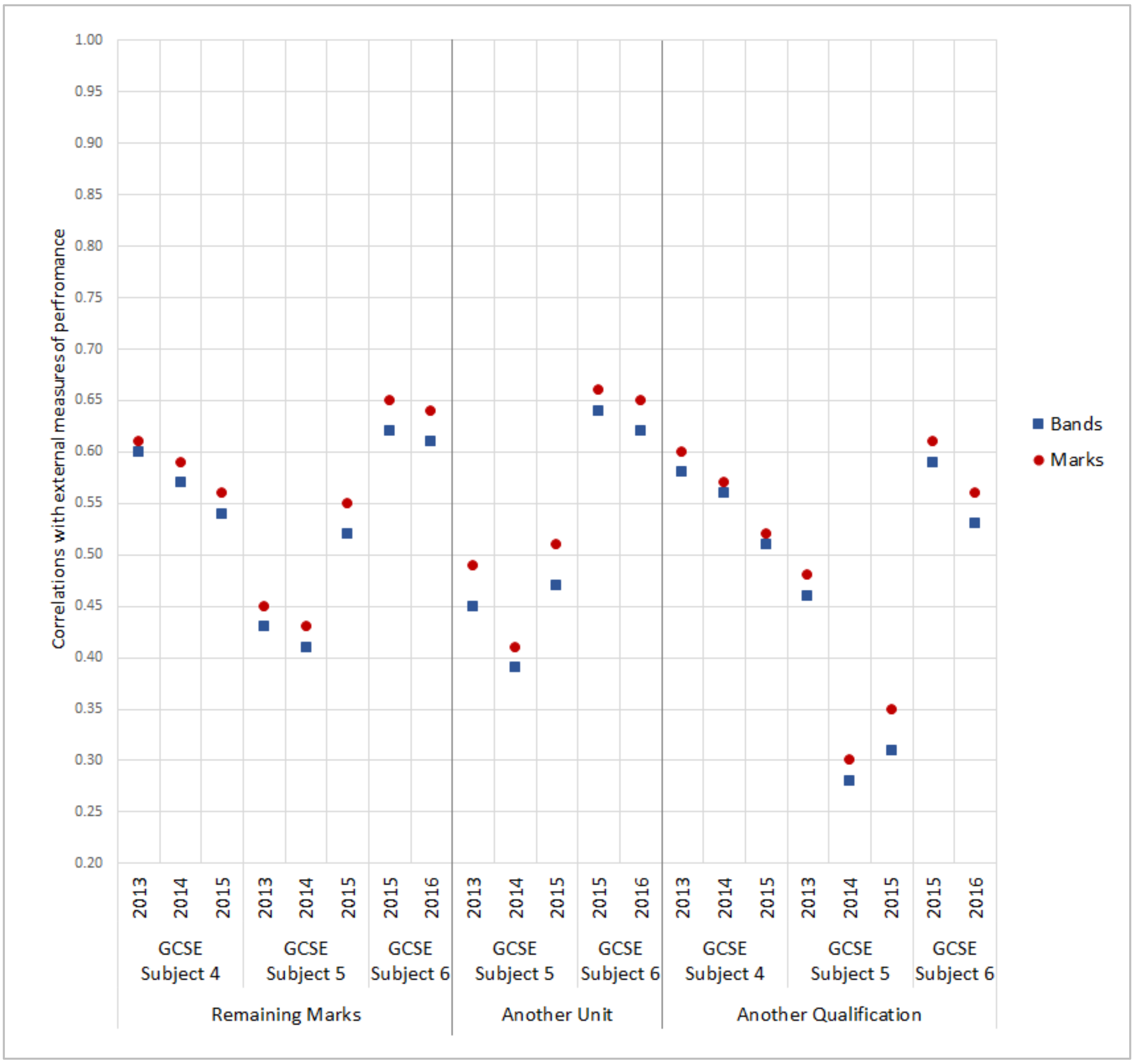

*Figure 2 The Spearman order rank correlation values for levels within the mark scheme (bands) and marks within the bands (marks) with external measures of performance for GCSE items explored.*

## Correlation of marks with external measures for each level within the mark scheme

To provide further detail on the value of marks within levels, for each qualification examined, correlations of detailed marks with selected external measure were calculated for each level within the mark scheme (see Tables 9 to 14). For brevity, only one external measure was included within this analysis: either the score on the other unit within the qualification or the performance on another qualification. Whichever measure had the greater number of available candidates was used. The vast majority of results revealed significant positive correlations between detailed marks and the external measure even when restricted to those marks within specific levels. As might be expected, statistical significance was achieved more often for those levels containing a large number of candidates (typically the middle levels).

*Table 9 The Spearman rank order correlations of marks within the level bands and another qualification measure for each level within the mark scheme for A level Subject 1 item.*

| A level Subject 1 item<br>with another unit measure | | | |
|---|---|---|---|
| Level | June 2013 | June 2014 | June 2015 |
| 5 | 0.29<br>p < .001<br>N = 221 | 0.29<br>p < .001<br>N = 410 | 0.21<br>p < .001<br>N = 409 |
| 4 | 0.18<br>p = .009<br>N = 215 | 0.11<br>p = .021<br>N = 458 | 0.16<br>p < .001<br>N = 508 |
| 3 | 0.14<br>p = .150<br>N = 105 | 0.15<br>p = .005<br>N = 345 | 0.15<br>p = .005<br>N = 360 |
| 2 | 0.52<br>p < .001<br>N = 43 | 0.12<br>p = .167<br>N = 144 | 0.14<br>p = .090<br>N = 145 |
| 1 | -0.23<br>p = .556<br>N = 9 | 0.29<br>p = .247<br>N = 18 | 0.25<br>p = .518<br>N = 9 |
| 0 | N/A | 0.32<br>p = .683<br>N = 4 | N/A |

*Table 10 The Spearman rank order correlations of marks within the level bands and another qualification measure for each level within the mark scheme for A level Subject 2 item.*

| A level Subject 2 item<br>with another unit measure | | | |
|---|---|---|---|
| Level | June 2013 | June 2014 | June 2015 |
| 4 | 0.10<br>p = .095<br>N = 300 | 0.16<br>p < .001<br>N = 770 | 0.03<br>p = .622<br>N = 339 |
| 3 | 0.24<br>p < .001<br>N = 2497 | 0.08<br>p = .463<br>N = 92 | 0.20<br>p < .001<br>N = 5277 |
| 2 | 0.12<br>p = .002<br>N = 716 | 0.08<br>p = .800<br>N = 12 | 0.13<br>p < .001<br>N = 1342 |
| 1 | 0.10<br>p = .365<br>N = 78 | N/A | -0.09<br>p = .337<br>N = 112 |
| 0 | N/A | N/A | N/A |

*Table 11 The Spearman rank order correlations of marks within the level bands and another qualification measure for each level within the mark scheme for A level Subject 3 item.*

| A level Subject 3 item<br>with another unit measure | | |
|---|---|---|
| Level | June 2014 | June 2015 |
| 6 | 0.09<br>p = .327<br>N = 125 | 0.06<br>p =.516<br>N = 122 |
| 5 | 0.07<br>p < .001<br>N = 1205 | 0.13<br>p < .001<br>N = 1177 |
| 4 | 0.15<br>p < .001<br>N = 3347 | 0.18<br>p < .001<br>N = 3298 |
| 3 | 0.19<br>p < .001<br>N = 2317 | 0.18<br>p < .001<br>N = 1925 |
| 2 | 0.05<br>p = .423<br>N = 250 | 0.07<br>p = .318<br>N = 217 |
| 1 | 0.16<br>p = .522<br>N = 18 | -0.01<br>p = .953<br>N = 23 |
| 0 | 0.14<br>p = .699<br>N = 10 | 0.56<br>p = .321<br>N = 5 |

*Table 12 The Spearman rank order correlations of marks within the level bands and another qualification measure for each level within the mark scheme for GCSE Subject 4 item.*

| GCSE Subject 4 item<br>with another qualification measure | | | |
|---|---|---|---|
| Level | June 2013 | June 2014 | June 2015 |
| 8 | 0.19<br>p < .001<br>N = 795 | 0.17<br>p < .001<br>N = 779 | 0.17<br>p < .001<br>N = 861 |
| 7 | 0.15<br>p < .001<br>N = 2274 | 0.19<br>p < .001<br>N = 2487 | 0.14<br>p < .001<br>N = 2484 |
| 6 | 0.18<br>p < .001<br>N = 4402 | 0.18<br>p < .001<br>N = 4175 | 0.14<br>p < .001<br>N = 3782 |
| 5 | 0.20 | 0.18 | 0.16 |

|  |  |  |  |
|---|---|---|---|
|  | p < .001<br>N = 3935 | p < .001<br>N = 3124 | p < .001<br>N = 2459 |
| 4 | 0.15<br>p < .001<br>N = 1293 | 0.16<br>p < .001<br>N = 770 | 0.16<br>p = .002<br>N = 635 |
| 3 | 0.16<br>p = .046<br>N = 162 | 0.08<br>p = .463<br>N = 92 | 0.02<br>p = .857<br>N = 72 |
| 2 | 0.24<br>p = .33<br>N = 18 | 0.08<br>p = .800<br>N = 12 | -0.28<br>p = .405<br>N = 11 |
| 1 | N/A | N/A | N/A |
| 0 | -0.29<br>p = .128<br>N = 28 | N/A | -0.41<br>p = .273<br>N = 9 |

***Table 13 The Spearman rank order correlations of marks within the level bands and another qualification measure for each level within the mark scheme for GCSE Subject 5 item.***

| GCSE Subject 5 item<br>with another unit measure |  |  |  |
|---|---|---|---|
| Level | June 2013 | June 2014 | June 2015 |
| 4 | 0.17<br>p = .002<br>N = 345 | 0.11<br>p < .001<br>N = 1047 | 0.12<br>p < .001<br>N = 1405 |
| 3 | 0.25<br>p < .001<br>N = 1956 | 0.17<br>p < .001<br>N = 5818 | 0.18<br>p < .001<br>N = 7810 |
| 2 | 0.26<br>p < .001<br>N = 2298 | 0.16<br>p < .001<br>N = 11684 | 0.26<br>p < .001<br>N = 11547 |
| 1 | 0.04<br>p = .403<br>N = 370 | 0.12<br>p < .001<br>N = 3805 | 0.25<br>p < .001<br>N = 1909 |
| 0 | N/A | N/A | N/A |

*Table 14 The Spearman rank order correlations of marks within the level bands and another qualification measure for each level within the mark scheme for GCSE Subject 6 item.*

| GCSE Subject 6 item<br>with another unit measure | | |
|---|---|---|
| Level | June 2015 | June 2016 |
| 6 | 0.17<br>p < .001<br>N = 542 | 0.22<br>p < .001<br>N = 241 |
| 5 | 0.25<br>p < .001<br>N = 2571 | 0.25<br>p < .001<br>N = 2026 |
| 4 | 0.27<br>p < .001<br>N = 4079 | 0.27<br>p < .001<br>N = 4481 |
| 3 | 0.28<br>p < .001<br>N = 3084 | 0.29<br>p < .001<br>N = 1844 |
| 2 | 0.28<br>p < .001<br>N = 1914 | 0.30<br>p < .001<br>N = 3750 |
| 1 | 0.20<br>p < .001<br>N = 1023 | 0.33<br>p < .001<br>N = 1905 |
| 0 | N/A | N/A |

## Discussion

The aim of this research was to evaluate the usefulness of detailed marks within levels-based mark schemes in comparison to the levels within the mark scheme alone. The results revealed using the detailed item marks within the levels leads to slightly higher predictive value than purely using performance captured by levels. This suggests that the detailed marks provide a more accurate picture of each candidate's performance. As such, to some extent, the findings from the current study support the use of detailed marks within the levels-based mark schemes.

Ahmed and Pollitt (2011) argued that making refined judgements about candidates' performance is more challenging than deciding on the level of response, as the difference between the single marks in terms of response quality is less pronounced than the difference between the levels within the mark scheme. Therefore, the benefit of using marks within the levels may come as a surprise. Nevertheless, the higher predictive value when using detailed marks indicates that rather than introducing unnecessary noise, the marks improve on the discriminatory value of the levels.

Although the research here relates to a physical aspect of the marking process, there is a clear relationship with more general earlier research on the optimal length of scoring scales. That is, using the levels within the mark scheme in place of the marks leads to a reduction of the original marking scale into the one with a smaller number of categories (i.e. there are fewer levels than there are marks). Bramley, Vidal Rodeiro and Wilson (2014) showed that condensing the scores on a long mark scale down into the equivalent positions on a shorter categorisation of that scale indeed results in a loss of information, albeit small. Therefore, from this point of view, smaller correlations are expected when a smaller number of categories are used in the analyses. The results obtained in this study are broadly consistent with expected losses in predictive value from shortening scales in general. In other words, although the way scales have been shortened in this study relates to a predefined part of the marking process (the levels and bands in the mark scheme), the loss of predictive value is consistent with an arbitrary process condensing the mark scale.

From both the current and previous research, it could be inferred that longer scales are better as they provide a more accurate picture of a candidate's performance. However, Shaw, Huffman and Haviland (1987) demonstrated that whereas increasing a number of intervals of a scale reduces information loss, eight categories retain 95% of the information of the original scale and the amount information gained with each additional category beyond this number is relatively small. Similarly, analysis by Bramley et al. (2014) revealed that recoding from a long scale with 169 categories to a shorter scale with just nine categories retains 97% of the information in the original scale. These studies show that information loss due to reduction of the scale into a relatively low number of categories is marginal as long as eight to nine categories are retained. The results from the current study seem to support this notion. For example, the differences in correlations of external measures with levels within the mark scheme and marks within the levels were in general smaller for GCSE Subject 4, with the nine levels within the mark scheme than for GCSE Subject 5, with only four levels within the mark scheme (see Figure 2).

One could argue that the informative value of marks and consistency between the markers could improve further if the definitions of the distinction among different marks within the mark scheme were clearly articulated (Ahmed & Pollitt, 2011, Shaw & Weir, 2007). The marks within a level are rarely described individually. Indeed, the mark schemes of the items explored in this study contained only descriptors of the responses needed to achieve a particular level. Including the descriptors of marks to better differentiate performance within a level could potentially increase the consistency between markers. However, even without these, the research in this report has indicated that detailed marks are of greater value than the pure levels (for which descriptors were available). This indicates that making detailed marks within levels available can be beneficial even without providing descriptors for each of them.

Increasing the amount of detail in mark schemes for extended response questions in a quest to improve reliability, would inadvertently introduce other issues. One may ask how precisely each mark can be defined without imposing excessive limitations on candidates' responses?

Constraining the mark scheme by specifying the required content of response and moving towards point-based mark scheme would undoubtedly threaten the validity of extended response questions, which inherently come with several valid approaches (Ahmed & Pollitt, 2011; Pinot de Moira, 2011).

Currently, there is no evidence about the reliability of levels-based mark schemes with a different number of levels/marks within each level, with mark schemes designed based on limits of cognitive discrimination of the responses and weight within the specification (Alpha Plus, 2014). The psychological literature on rating scales suggests that the optimal number of response categories in terms of reliability, validity and degree of differentiation could be achieved with approximately five to seven categories (Dawes, 2008, Givon & Shapira, 1984; Lozano, García-Cueto, & Muñiz, 2008). However, it is not clear how this research applies to the context of marking. There is some indication that the number of levels within the mark scheme should be more important than the width of the level (Shaw et al., 1987). However, previous work has also suggested that marking is more reliable when each band is composed of the same number of marks (Pinot de Moira, 2013).[1] Nevertheless, further research is required to determine the optimal ratio between the number of levels within the mark scheme and number of marks within each level. For example, differences in the predictive value of levels could be explored by manipulating the fine-grained marks obtained by the candidates, which could be broken into different levels while keeping the approximate distribution of proportions of candidates scoring in each level.

## Conclusion

The findings from the current study support the use of detailed marks within the levels-based mark schemes, which generally provide a significant improvement over using the levels within the mark scheme alone. In some ways, this result is unsurprising as condensing a marking scale to one with a lower number of categories is expected lead to some loss of information.

The benefit of using detailed marks within the mark scheme is likely to depend on the quality of the mark scheme itself. Previous research has suggested that mark schemes containing fine-grained descriptions that help markers to assign an appropriate mark to a candidate's response within each level are likely to increase markers' agreement but will consequentially reduce the marking speed, as it will take longer to apply them. Future research is required to establish the optimal ratio of levels within the mark scheme and marks within each level.

## References

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

[1] For the items explored in the current study, there was no clear pattern as to whether having equally wide levels resulted in less pronounced differences between the accuracy of levels versus marks.

AlphaPlus. (2014). *Standardisation methods, mark schemes, and their impact on marking reliability.*

Benton, T., & Gallacher, T. (2018). Is comparative judgement just a quick form of multiple marking. *Research Matters: A Cambridge Assessment Publication (26)*, 22-28.

Black, B., Suto, I., & 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 and Practice, 18*(3), 295-318.

Bramley, T. (2009). Mark scheme features associated with different levels of marker agreement. *Research Matters: A Cambridge Assessment Publication* (8), 16-23.

Bramley, T., Vidal Rodeiro, C. & Wilson, F. (2014). *Reporting scale scores at GCSE and A level.* Cambridge Assessment Internal Research Report. Cambridge, UK: Cambridge Assessment.

Child, S., Munro, J., & Benton, T. (2015). *An experimental investigation of the effects of mark scheme features on marking reliability*. Cambridge Assessment External Research Report. Cambridge, UK: Cambridge Assessment. Available at: https://www.cambridgeassessment.org.uk/Images/417277-an-experimental-investigation-of-the-effects-of-mark-scheme-features-on-marking-reliability.pdf

Dawes, J. G. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point, and 10-point scales. *International Journal of Market Research*, *50*, 61-77.

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

Givon, M. M., & Shapira, Z. (1984). Response to rating scales: A theoretical model and its application to the number of categories problem. *Journal of Marketing Research, 21*, 410-419.

Hughes, S. R. (2018). Understanding the impact of levels based mark scheme design. Paper presented at the Ofqual Seminar, Coventry, UK.

Hughes, S., & Shaw, S. (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. *Research Matters: A Cambridge Assessment Publication* (21), 42-48.

Lee, I. A., & Preacher, K. J. (2013, September). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]. Available from http://quantpsy.org.

Lozano, L. M., García-Cueto, E. M., Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 4, 73-79.

Massey, A.J., & Raikes, N. (2006). Item-level examiner agreement. Paper presented at the annual conference of the British Educational Research Association, Warwick, UK.

Ofqual. (2011). *GCSE, GCE, Principal learning and project code of practice.* Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/371268/2011-05-27-code-of-practice.pdf

Ofqual. (2014). *Quality of Marking: Review of Literature on Item-level Marking Research.* Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/605659/2014-02-14-quality-of-marking-review-of-literature-on-item-level-marking-research.pdf.

Pinot de Moira, A. (2011). *Levels-based mark schemes and marking bias*. Manchester: AQA Centre for Education Research and Practice.

Pinot de Moira, A. (2013). *Features of a levels-based mark scheme and their effect on marking reliability.* Available at: https://cerp.aqa.org.uk/sites/default/files/pdf_upload/CERP_TR_APM_03042013.pdf

Shaw, D. G., Huffman, M. D., & Haviland, M. G. (1987). Grouping continuous data in discrete intervals: information loss and recovery. *Journal of Educational Measurement, 24*(2), 167-173.

Shaw, S. D. & 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.

Suto, W. M. I., & Greatorex, J. (2008). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policy & Practice, 15*(1), 73-89.

Suto, I., & Nadas, R. (2008). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education, 23*(4), 477-497.

Tisi, J., Whitehouse, G., Maughan, S. & Burdett, N. (2013). *A Review of literature on marking reliability research* (report for Ofqual). Slough, NFER.