Cambridge Assessment is the brand name of the University of Cambridge Local Examinations Syndicate, a department of the University of Cambridge. Cambridge Assessment is a not-for-profit organisation.

**Citation**

Research Matters : 14
A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

One of the major problems in educational research is the time lag involved in evaluation. An innovation is conceived, design work is completed, implementation is undertaken, and evaluation starts. But evaluation takes time. It is important to follow a cohort completely through the programme and the assessments linked to a new qualification. This typically will take two years (GCSE and full A level) – with additional time for analysis and writing up. This raises significant ethical issues and pressure for wide implementation of new qualifications can be in tension with the need for ethically-based piloting. Alongside this, policy changes at national level occur with increasing frequency, often before the evaluation of the previous form of the qualification has reported. Despite the increasing problem of evaluations reporting after the point at which the qualification to which they relate has been significantly changed or superceded, Cambridge Assessment believes that we should continue to have an ethically-based commitment to timely and incisive evaluation, and that we should work to reduce the frequency of unnecessary fundamental change to the form and content of qualifications. Time lag means that good judgement needs to be exercised in commissioning evaluations – they need to be planned and put in place in full recognition that evaluation and reporting takes time. Hence in this issue, a key article is Vidal Rodeiro’s examination of the impact of GCSE modularisation. The rush towards widespread modularisation of GCSE has been replaced by rapid reversal towards linear GCSEs, but it is vital to understand, in detail, the impact of that drive to modularisation. Modularisation has not been free of problems in the past – witness the so-called ‘crisis’ of 2002 – and it is essential to understand fully both emergent and deeper problems and to examine critically the balance of benefits and unintended consequences. If such evaluations decline, genuine scientific accumulation of knowledge regarding assessment and qualifications comes under threat, and unfounded pendulum swings in policy are likely to increase in frequency and intensity.

Tim Oates Group Director, Assessment Research and Development

Editorial

Most of the articles in this issue address matters relating to comparability covering a range of contexts from methodology to modularisation and highlighting challenges posed in this contentious field. In the first article Yim reports on research carried out to investigate a rank ordering method using pseudo candidates’ scripts and real candidates’ scripts in a rank-ordering exercise at syllabus level. His research investigated a fundamental question where teachers allocate candidates to two different areas of demands analysis in the context of comparisons involving general and vocational qualifications. Most of the articles in this issue address matters relating to comparability covering a range of contexts from methodology to modularisation and highlighting challenges posed in this contentious field. In the first article Yim reports on research carried out to investigate a rank ordering method using pseudo candidates’ scripts and real candidates’ scripts. The aim was to explore, qualitatively and quantitatively, the impact of comparing different script types, particularly in relation to different response styles. Rushton focuses on holistic judgements where candidates’ performances are uneven to investigate whether uneven profiles make the judge’s task more difficult. Both of these articles add to the methodological debate about how best to design and conduct comparability studies which involve human judgement.

Dhawan reports on his work on comparable grade boundaries across tiered examination papers. His research investigated a fundamental question where teachers allocate candidates to two different versions of a test based on their ability. The findings from this study were encouraging and potential methods for future routine analyses were outlined. In her work on the impact of modularisation Rodeiro has investigated a range of issues and in this article she focuses on whether different assessment routes (linear vs modular) equip students equally for further study.

The next comparability challenge addressed in this issue is, perhaps, the most complex. It combines the area of demands analysis in the context of comparisons involving general and vocational qualifications. Greatorex and Shiell report on a pilot study of a method for use in comparability research which was designed to ascertain whether the demands instrument used was appropriate as a tool in research studies and to explore how judges made decisions in this context. Greatorex and Shiell’s article reports on research into validity in the context of teacher assessed Independent Research Reports that contribute to the Cambridge Pre-U Global Perspectives and Research. This study forms part of a wider programme of research into the IRR.

In their article on bilingualism Imam and Shaw add their voices to a complex debate. They discuss the educational context within which students prepare for Cambridge International Examinations assessments and they address a number of key issues.

In the final contribution Emery, Gill, Grayson and Vidal outline highlights from the Cambridge Assessment Statistics Reports Series. These reports cover various aspects of the examinations system and are produced each year based on the latest national examinations data.

Sylvia Green Director of Research
An intra-board comparison of the effect of using pseudo candidates’ scripts and real candidates’ scripts in a rank-ordering exercise at syllabus level

Louis Yim Cambridge International Examinations

1. Introduction

There are a number of examination boards offering public examinations in England which lead to the same qualifications, for example GCSE and GCE A level. Although each examination syllabus must conform to general qualifications criteria approved by the examination regulator, and also to a common core of subject content, the syllabuses may differ between boards in other respects. A crucial question of whether it is easier to obtain a given grade in a particular examination with one board than with another arises. In fact, this issue is not limited to England alone, but extends to overseas countries where candidates sit for examinations which are claimed to be equivalent qualifications to the GCSE and GCE A level.

To ensure the equivalence of standards of similar qualifications across different examination boards, several research programmes have been conducted, most of which only compare examination standards qualitatively between examination boards by reporting, say, ‘Board X is harder (or easier) than Board Y’ without quantifying the difference in standards. A rank-ordering method is a recent addition to a wide selection of comparability methodologies which has been used relatively effectively to compare standards quantitatively across examination boards at component level (Bramley, 2005; Bramley, 2007), as well as at syllabus level within the same subject (Yim and Shaw, 2009).

The notion of a pseudo (or composite) candidate has been adopted for syllabus level comparability exercises. A candidate is a composition of different candidates sitting the same examination from the same examination board. The phrase ‘scripts of a pseudo candidate’ at syllabus level effectively means scripts of prescribed components with specific marks, contributing to the grading of a particular assessment, from different candidates which mimic the profile of component marks of an ordinary (or real) candidate. Although other studies/literature have briefly mentioned the probable impact of using pseudo candidates’ scripts in comparability studies, that is, that judges/examiners would find them harder to assess (Irlett, 2002; Guthrie, 2003; Bramley, 2007), the claim could not be substantiated until a recent comparability study on the effect of using different types of candidates’ scripts (pseudo and real candidates) had been carried out (Yim and Forster, 2010). That study showed that the use of different types of candidates’ scripts (pseudo and real candidates) by expert judges at syllabus level during a comparability exercise would have an effect on judges’ decisions on candidates’ performance. This could primarily be accounted for by a disparity of response style in each component in pseudo candidates’ scripts; whereas there is no apparent disparity in real candidates’ scripts.

The study reported here attempted to further refine the design of the Yim and Forster (ibid.) study in order to focus solely on the contrast between scripts of pseudo and real candidates. Instead of comparing scripts from two different examination boards (where the syllabus content and assessment structure can differ slightly), in this study two parallel assessments of the same syllabus from the same board were compared.

As a further control, an assessment with a large examination cohort was chosen, so that scripts from the real and pseudo candidates could be selected or created respectively such that they had very similar profiles of marks (scores) across the components of the assessment. For further research on the effect of mark profile on expert judgement of script quality, see Rushton (this issue, p.10).

The rationale behind conducting research at syllabus level is that quantitative results can generally help inform CIE’s grading decisions in terms of grade boundary adjustment at component level for the assessment of a particular syllabus. The materials used in this study were question papers, mark schemes, syllabus specifications and two types of scripts (pseudo and real candidates) for all components from the same examination session within the same examination board. These were then evaluated by expert judges to generate rankings in terms of perceived quality of both pseudo and real candidates’ scripts. The resulting data were analysed using the multifacet Rasch modelling technique (Linacre, 1987) and the difference in standards between pseudo and real candidates’ scripts was deduced from graphs. The methodology, the research outcome, and judges’ feedback are described below.

Background to comparability exercises

Comparability in this context is concerned with the application of the same standard across different examinations (Newton, 2007). The purpose of inter-board comparability studies is to compare standards across different examination boards. In making this comparison, it is important to distinguish between content standards and performance standards: “Content standards refer to the curriculum (or syllabus/specification) and what examinees are expected to know and to be able to do … performance standards communicate how well examinees are expected to perform in relation to the content standards” (Hambleton, 2001). In fact, a more precise definition of

1. The Office of Qualifications and Examinations Regulation (Ofqual), England
2. In Cambridge International Examinations (CIE), the assessment of the full syllabus usually comprises several different components, for example two written examination papers and a practical examination.
3. External consultants with subject matter expertise. Usually they are or have been senior examiners in the subject.
comparability is paramount since qualifications can be compared on many different aspects, such as demand of the curriculum, similarity of content materials, difficulty experienced by candidates, demand of assessment materials, perceived quality of candidate outcome based on scripts and standards of attainment, etc.

One way to compare performance standards across assessments from different boards (or across parallel assessments from the same board) is to ask experts to compare pairs of scripts from each assessment and make judgements about which one demonstrates better quality. Such exercises address the question: “Which syllabuses’ grade boundary scripts are perceived by expert judges to be of better quality (after allowing for slight difference of syllabus content, question paper and mark scheme difficulty)?”

One way of analysing the data from these paired comparison judgements is by Thurstone’s model (case 5) for comparative judgements (Thurstone, 1927). For a discussion of how Thurstone’s method has been applied in the context of examination comparability, see Bramley (2007). For recent applications of the method see Yim, Shaw and Lewis (2008), and Yim and Shaw (2009).

The main advantage of this approach is that the use of candidates’ scripts provides explicit evidence of the knowledge, understanding and skills of examinees, and hence direct comparison of performance standards can be achieved. For inter-board comparisons it should be noted that it is only possible to compare performance standards if the content standards across the examination boards are similar enough for the different assessments to be considered to be measuring the same construct (underlying trait). If the question papers, mark schemes and syllabus specifications are very different, examiners will be expected to make judgements about the relative performance standards in a context of possible differences in content standards. The outcome of such an exercise would be rendered less reliable as a result of disparate schemes of assessment and syllabus contents.

In practice, the nature of the scripts (objects) being compared is such that the scripts take a long time to read, and paired comparisons are unlikely to be independent because of the repeated use of shared scripts. Hence examiners might already have the knowledge of either or both of the scripts before the paired comparisons which violates the assumption of local independence between paired judgements. Therefore instead of asking judges to make paired comparisons, it is less time-consuming to ask them to put sets of scripts into rank-order of perceived quality. It is then possible to extract paired comparison data from the rank-order in the form of ‘1 beats 2’, ‘2 beats 3’, ‘1 beats 3’ and so on (Bramley, 2007). These extracted paired comparisons are not statistically independent, because they are constrained by the ranking, but as explained above even genuine paired judgements would arguably not be independent either. In other words, a rank-ordering method is a time-saving variant of the paired comparison method for comparing performance standards. Such comparison exercises draw heavily on the expertise of senior examiners to judge the quality of examinees’ work, taking into account the demand placed upon examinees by the individual syllabuses/specifications, question papers and mark schemes.

Rationale behind using pseudo candidates’ scripts versus real candidates’ scripts

The rank-ordering method at syllabus level using pseudo candidates’ scripts has demonstrated that the use of careful pack design of scripts and a multifacet Rasch modelling technique can yield a quantitative difference in standards between two examination boards, which can inform grade boundary adjustment during awarding meetings if there is a need to align standards with another exam board (Yim, Shaw and Lewis, 2008; Yim and Shaw, 2009). The rationale behind using pseudo candidates’ scripts instead of real candidates’ scripts is to provide examiners with an exact ‘flat’ profile of candidates’ performance at component level for a particular syllabus grade since real candidates with an exact ‘flat’ profile are rare. A candidate with an exact ‘flat’ profile on a three-component assessment could be considered to be one who achieves a mark exactly at the grade boundary of, say, B at syllabus level with all three components also being at a mark exactly at the grade boundary of B; a candidate with an uneven profile could be considered as one achieves a mark at the grade boundary of B at syllabus level, but with uneven grades at component level, for example, a mark at well above grade A in Component 1, a mark at the boundary of grade B in Component 2 and a mark at the middle of grade C in Component 3. The latter is more common/authentic in examination practice. The use of the exact ‘flat’ profile is to indicate to examiners that a clear-cut standard across component level, for example, all components at the boundary of grade B, will lead to the same syllabus grade level, that is, grade B. This is intended to facilitate the judgement process of rank-ordering for examiners. Although examiners have been able to complete their judgements with merely slight difficulties, some of the qualitative feedback received in previous studies suggested that the use of real candidates’ scripts could minimise a change of style in candidates’ response between different components and hence that examiners would be more confident on their rank-ordering results, albeit sacrificing the exact ‘flat’ performance profile. The purpose of this study was to compare the results of rank-ordering scripts from real and pseudo candidates with the same mark profiles across the components, so that any differences in outcome or in the examiners’ reported experience would relate only to the pseudo/real distinction and not to the mark profiles of the scripts they compared.

2. Method

This study used the same procedures as the previous study (Yim and Forster, 2010) in terms of the algorithm for selecting pseudo and real candidates, the pack design, the instructions given to the expert judges, and the data analysis method. The only difference was that the syllabus comparison was within the same examination board, that is, an intra-board comparison instead of an inter-board comparison, in an attempt to focus solely on the effect of using pseudo and real candidates’ scripts in the rank-ordering method at syllabus level.

The materials required in this project were question papers, mark schemes, syllabus specification and candidates’ scripts (both pseudo and real candidates) from the examination board. Seventeen exact ‘flat’ profiles of pseudo candidates’ scripts at grade boundaries, A, B, C, D and E, and their intermediate grade boundaries at 2/3 and 1/3 of a grade above each grade, and 1/3 and 2/3 of a grade below each grade for both

---

4. At awarding meetings the grade boundary locations on the raw mark scale of each component are decided.

5. There is a subtle difference between a candidate with an exact even (or ‘flat’) profile and one with an even profile in this discussion. The criteria of the former are a candidate with the targeted component marks at exactly the same point relative to the grade boundary, whereas the latter only requires the same grades across prescribed components (e.g. BBB) within a syllabus and no stipulation of any targeted component marks.
assessments were selected. The first assessment is referred to as ‘Option AA’ and the second as ‘Option BB’ in this article.

Instead of using random real candidates’ scripts at particular syllabus marks/grade levels, real candidates were selected whose script component marks fit within a 1% of each targeted component mark of their pseudo candidates’ counterparts. The intention was to ensure both pseudo and real candidates’ scripts had the exact ‘flat’ profile to rule out differences in component marks as a potential feature influencing the comparison. It should be noted that the selection of real candidates’ scripts meeting this criterion can only work well in examinations with a large entry as there are more scripts to choose from.

After selecting the pseudo and real candidates’ scripts, examiner markings/annotations were removed electronically via a scanner such that they did not have an influence on the rank-ordering judgements during the experts’ judging process. Each candidate (pseudo and real) was then allocated into different pack of scripts in accordance with the pack design.

Each pack comprised six candidates (three from Option AA and three from Option BB) and there were altogether eight packs (A to E) for each type of script, that is, pseudo and real. The candidates and hence their scripts in each pack were randomised, coded and labelled such that the original scripts’ rank-order based on marks was concealed. Each candidate’s scripts were photocopied for each expert judge.

In each pack of six scripts, two were common to the pack above and two were common to the pack below (where ‘above’ and ‘below’ refer to the rank order by total mark). The top pack had two scripts in common with the pack below and the bottom pack had two scripts in common with the pack above. This linked design allowed a common scale of ‘perceived quality’ to be created from the ranking judgements.

Five senior examiners (expert judges), all with marking/moderating experience of the syllabus concerned, were recruited to make judgements about both pseudo and real candidates’ scripts in two phases. In phase I, three expert judges were allocated pseudo candidates’ scripts and two were allocated real candidates’ scripts; in phase II, the nature of the scripts was swapped such that each expert judge had judged both the pseudo and real candidates’ scripts at the completion of the study. This was to cancel out any effect due to the script-judging order. The gap between the two phases was two weeks, with the same judges participating in both phases. Their task was to rank-order scripts within each pack from best (highest quality = 1) to worst (lowest quality = 6) based on a holistic judgement and record their outcomes in the tables provided on a record sheet.

There was a gap of two weeks between the two phases to ensure that there was no cross-over of judges’ rank-ordering experience. Each expert judge was asked to complete a questionnaire towards the end of each phase for the qualitative analysis of the study.

3. Analysis and results

Once the rank-order data were received from examiners, they were deconstructed into paired comparison data and then analysed using the Rasch analysis (FACETS) software (Linacre, 1987) to estimate the difficulty/ability of each script/candidate based on the inter-relationship of examiners’ rankings. A one-facet model was used which estimated a measure of ‘perceived quality’ (‘Measure’) for each script in the study.

Extracts from the FACETS output are given in Appendix A.

The separation reliability index (analogous to Cronbach’s Alpha) was high in both types of scripts, that is, 0.98, showing that the differences in perceived quality among the scripts could not be attributed to chance. There are different views on what fit index is actually acceptable, however, based on operational experience the lower and upper limits of 0.7 and 1.6 respectively for mean squares seems to be useful and acceptable for practical purposes and were used in this analysis. The fit statistics from the infit and outfit columns of the FACET output for scripts and judges in both real and pseudo cases showed that the data were predicted well by the Rasch model. All these scale statistics need to be treated with caution because, as mentioned previously, the paired comparison analysis violates the assumption of local independence between paired judgements when derived from the rank-ordering outcome.

Figures 1 and 2 show the results of the comparability plots for the pseudo and real candidates respectively. The vertical axis along the left of the figures represents the ‘Measure’ (or script quality) scale in log-odds units (logits). A distance of 1.1 logits corresponds to a probability of 75% that the script with the higher measure will be ranked above a script with the lower measure. In these graphs each data point (diamond – Option AA and square – Option BB) represents a script. Each script (a data point) is positioned according to its measure. Thus performances are rank ordered with the most able candidates at the top of the axis and the least able at the bottom, that is, the scripts in the top half of the graph (above 0 logits) are judged to be of better quality than those in the bottom half (below 0 logits). The horizontal axis shows the overall syllabus aggregate percentage mark obtained from conventional marking of the scripts.

The two straight lines in each comparability plot shown in Figures 1 and 2 are linear regression lines whose equations are given in the boxes. The parameter R is the correlation coefficient. The magnitude of R indicates the extent to which the two sets of measurements (‘Measure’ and ‘Syllabus %’) are linearly related. The pair of regression lines, that is, Options AA and BB, in the pseudo and real candidates’ cases) shares similar features such as strong correlation, similar gradient, no reversal of position; that is, option AA regression line is consistently on top of Option BB. Tables 1 and 2 show the comparison of some numerical findings between the pseudo and real candidates’ cases.

Table 1: Differences in ‘Measure’ (along the y-axis) between Option AA and Option BB at Grades A, B, C, D and E for both pseudo and real candidates’ cases

<table>
<thead>
<tr>
<th>Types of scripts</th>
<th>( \Delta_{\text{measure}} ) [logits]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Pseudo candidates</td>
<td>0.49</td>
</tr>
<tr>
<td>Real candidates</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Table 1 tabulates the differences in ‘Measure’ (along the y-axis) between Option AA and Option BB at Grades A, B, C, D and E for both pseudo and real candidates’ cases. In an ideal case the values of \( \Delta_{\text{measure}} \) as shown in Figure 2, in both pseudo and real candidates’ cases should be in line with one another, but the differences in Table 1 suggest that there are disparities at all grades, albeit small. In other words, the recommendations for grade boundary adjustments at syllabus level to achieve the equivalence of standards between options are different.
Figure 1: A comparability plot from grades A to E for pseudo candidates’ scripts between Options AA and BB

Figure 2: A comparability plot from grades A to E for real candidates’ scripts between Options AA and BB
depending on the type of candidates’ scripts being used. This outcome could be explained by the fact that examiners were using a completely different set of scripts but with almost identical component marks contributing to the same syllabus marks/grade level based on a careful script selection for the two evaluation phases. The small differences between the pseudo and real candidates’ cases at each grade are, in fact, rather encouraging as they demonstrate that the rank-ordering method could, to a certain extent, produce similar results.

Table 2: A comparison of the correlation coefficient $R$ between ‘Measure’ and ‘Syllabus %’ for the pseudo and real candidates’ cases

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Type of scripts</th>
<th>Correlation coefficient ($R$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>Pseudo candidates</td>
<td>0.948</td>
</tr>
<tr>
<td></td>
<td>Real candidates</td>
<td>0.945</td>
</tr>
<tr>
<td>BB</td>
<td>Pseudo candidates</td>
<td>0.979</td>
</tr>
<tr>
<td></td>
<td>Real candidates</td>
<td>0.942</td>
</tr>
</tbody>
</table>

Table 2 shows a comparison of the correlation coefficient ($R$) between ‘Measure’ and ‘Syllabus %’ for the pseudo and real candidates in Options AA and BB. The correlations for pseudo and real candidates’ cases were very similar within the same assessment and across assessments. The strong correlations ($R \geq 0.942$) in all cases between the ‘Measure’ and the ‘Syllabus %’ show that the trait of holistic quality as perceived by the judges was very similar to the trait of quality as rewarded by the mark scheme. The correlations were either the same or fractionally higher in the pseudo candidates’ cases, a finding that departs from those obtained from the previous inter-board comparability study (Yim and Forster, 2010) where the correlation in the pseudo candidates’ cases were consistently lower than those of the real candidates’ case. It should be recalled that the only difference in terms of the research design between the previous comparability study and the current one was that in this study, both assessments were from the same syllabus from the same examination board.

4. Feedback from examiners

Responses on questionnaires were collected from five examiners who carried out both phases to help understand the qualitative aspects of their rank-ordering experience relating to the overall difficulty of the task, the amount of time taken to rank order the scripts, what made some packs more or less difficult to rank, the difficulties presented by different types of scripts, any differences in the task between papers, and the strategy they deployed.

Overall difficulty of the task

All five participants were senior examiners and had taken part in at least one rank-ordering exercise previously. Four of them found the task ‘fairly difficult’ to execute; and one examiner found it ‘fairly easy’. Reasons for difficulty are tabulated in Table 3.

Examiners tended to take between 30 and 80 minutes per pack during the evaluation for pseudo candidates’ scripts and between 30 and 90 minutes per pack for real candidates’ scripts. Four out of five examiners did not think the length of time for the evaluation varied much from pack to pack.

Differences were also reported relating to the ease or difficulty of rank-ordering certain packs. Scripts from more able candidates were the most time-consuming to rank order although they were slightly less problematic as there was perceived to be a wider range of ability instantiated in performances. Scripts from less able candidates were more difficult to rank, and standards were perceived to be more closely grouped. Other factors included topic variation (student strengths being topic-related).

Four out of five examiners after completing the two-phase evaluation suggested that the task of rank-ordering real candidates’ scripts was easier or much easier. Examiners articulated a range of difficulties associated with the nature of the pseudo candidate profile:

- Pseudo candidates invariably demonstrate differing strengths; whereas real candidates might give more clues along the way (a reason also given in Arlett, 2002, and Guthrie, 2003).
- An inauthentic performance profile makes it difficult to develop an overview of candidates’ ability.
- Pseudo candidates’ scripts give evidence of different pedagogical heritage.

One examiner felt that both phases were of equal difficulty. He was surprised to find that the task had not been made easier by using real candidates’ scripts because of the amount of script information he needed to ‘keep in mind’ in order to carry out the judging.

Three out of five examiners felt that it was possible to carry out the judging for a pack of six candidates at syllabus level with three component papers. All examiners agreed that the task would have been made easier if they rank ordered individual scripts at component level instead.

Rank-ordering strategy

Examiners were allowed to adopt their own rank-ordering strategy during the evaluation phase though they were not allowed to re-mark the scripts. A variety of strategies were identified as follows:

- Identification of questions attempted by less able students: based on examiners’ experience, some questions can act as an indicator to distinguish between able and less able candidates.
- Use multiple choice paper to generalise candidates’ knowledge/understanding; this is followed by reviewing the written papers in depth for fine tuning candidates’ rank order.
- Identification of common and indicative questions across question papers to evaluate candidates’ ability.
- Overall judgement of depth and accuracy of answers.
Only a few examiners indicated a change of approach as the rank order task became increasingly more familiar. With experience, greater confidence was placed in subsequent judgements; and a greater tendency to revisit and overturn earlier judgements was also reported (as also found by Jones, Meadows and Al-Bayatti, 2004).

Examiners were uncertain as to whether more or less time on each script made any difference to the final rank order. However, in the main, they believed that a reduction or extension in the time taken to undertake the exercise would have little impact on the outcome.

5. Conclusions

The results of the comparison showed that the recommendations for grade boundary adjustments at syllabus level to achieve the equivalence of standards between exam boards were different depending on the type of candidates’ scripts being used, but that these differences were fairly small. This outcome could be explained by the fact that examiners were using completely different sets of scripts but with almost identical component marks contributing to the same syllabus marks/grade level based on a careful script selection for the two evaluation phases. The small differences between the pseudo and real candidates’ cases at each grade boundary are, in fact, rather encouraging as they demonstrate that the rank-ordering method could, to a certain extent, produce comparable results when conducted repeatedly.

In the current study the correlations between perceived quality and aggregate mark were consistently high and similar across the different conditions. In fact, they were consistently slightly higher in the pseudo candidates’ case. The implication of this finding is that the use of different types of candidates’ scripts does not affect how the trait of holistic quality is perceived, which departs from the previous findings (Yim and Forster, 2010) which suggested that the use of real candidates’ scripts could improve the correlation.

Although the prima facie evidence of the current study suggests that there is no preference in terms of using either type of scripts in terms of the internal quality of the scale produced (separation reliability and fit), or its correlation with an external variable (aggregate Syllabus % mark), the qualitative feedback from almost all expert judges suggests that the rank-ordering task had been made easier or much easier by using real candidates’ scripts. They felt more confident in carrying out the tasks as well as their rank-order outcomes. An in-depth comparison of the research outcome between inter-board and intra-board comparability studies, and the use of component level rank ordering methodology to infer outcome at syllabus level will constitute areas for further research.

References

Arlett, S (2002). A Study in VCE Health and Social Care, Units 1, 2 and 5. A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2001 examination and organised by AQA on behalf of the Joint Council for General Qualifications.


### Table 7.1.1 Judge Measurement Report (arranged by mN)

<table>
<thead>
<tr>
<th>Score</th>
<th>Count</th>
<th>Average</th>
<th>Average</th>
<th>Measure</th>
<th>S.E.</th>
<th>MnSq ZStd</th>
<th>MnSq ZStd</th>
<th>Discrn</th>
<th>PtMea</th>
<th>PtExp</th>
<th>Nu Script</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>Total</td>
<td>Obsvd</td>
<td>Fair-M</td>
<td>Score</td>
<td>Count</td>
<td>Average</td>
<td>Average</td>
<td>Measure</td>
<td>S.E.</td>
<td>MnSq ZStd</td>
<td>MnSq ZStd</td>
</tr>
<tr>
<td>60</td>
<td>120</td>
<td>.5</td>
<td>.50</td>
<td>.00</td>
<td>.21</td>
<td>1.20</td>
<td>2.0</td>
<td>1.20</td>
<td>1.11</td>
<td>.59</td>
<td>.39</td>
</tr>
<tr>
<td>60</td>
<td>120</td>
<td>.5</td>
<td>.50</td>
<td>.00</td>
<td>.21</td>
<td>1.01</td>
<td>.1</td>
<td>1.14</td>
<td>.8</td>
<td>.95</td>
<td>.49</td>
</tr>
<tr>
<td>60</td>
<td>120</td>
<td>.5</td>
<td>.50</td>
<td>.00</td>
<td>.21</td>
<td>1.01</td>
<td>.1</td>
<td>1.01</td>
<td>.1</td>
<td>.97</td>
<td>.50</td>
</tr>
<tr>
<td>60</td>
<td>120</td>
<td>.5</td>
<td>.50</td>
<td>.00</td>
<td>.21</td>
<td>.79</td>
<td>2.5</td>
<td>.69</td>
<td>2.0</td>
<td>1.47</td>
<td>.64</td>
</tr>
<tr>
<td>60</td>
<td>120</td>
<td>.5</td>
<td>.50</td>
<td>.00</td>
<td>.21</td>
<td>.95</td>
<td>.5</td>
<td>.89</td>
<td>-.6</td>
<td>1.12</td>
<td>.54</td>
</tr>
<tr>
<td>60.0</td>
<td>120.0</td>
<td>.5</td>
<td>.50</td>
<td>.00</td>
<td>.21</td>
<td>.95</td>
<td>.5</td>
<td>.89</td>
<td>-.6</td>
<td>1.12</td>
<td>.54</td>
</tr>
<tr>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>1.3</td>
<td>1.5</td>
<td>1.8</td>
<td>1.2</td>
<td>.08</td>
<td>.08</td>
</tr>
<tr>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>1.5</td>
<td>1.6</td>
<td>2.0</td>
<td>1.3</td>
<td>.09</td>
<td>.09</td>
</tr>
</tbody>
</table>

Model, Populn: RMSE .21 Adj (True) S.D. .00 Separation .00 Reliability 1.00
Model, Sample: RMSE .21 Adj (True) S.D. .00 Separation .00 Reliability .80
Model, Fixed (all same) chi-square: .0 d.f.: 4 significance (probability): 1.00

### Table 7.1.2 Script Measurement Report (arranged by mN)

<table>
<thead>
<tr>
<th>Score</th>
<th>Count</th>
<th>Average</th>
<th>Average</th>
<th>Measure</th>
<th>S.E.</th>
<th>MnSq ZStd</th>
<th>MnSq ZStd</th>
<th>Discrn</th>
<th>PtMea</th>
<th>PtExp</th>
<th>Nu Script</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>Total</td>
<td>Obsvd</td>
<td>Fair-M</td>
<td>Score</td>
<td>Count</td>
<td>Average</td>
<td>Average</td>
<td>Measure</td>
<td>S.E.</td>
<td>MnSq ZStd</td>
<td>MnSq ZStd</td>
</tr>
<tr>
<td>60</td>
<td>120</td>
<td>.5</td>
<td>.50</td>
<td>.00</td>
<td>.21</td>
<td>1.20</td>
<td>2.0</td>
<td>1.20</td>
<td>1.11</td>
<td>.59</td>
<td>.39</td>
</tr>
<tr>
<td>60</td>
<td>120</td>
<td>.5</td>
<td>.50</td>
<td>.00</td>
<td>.21</td>
<td>1.01</td>
<td>.1</td>
<td>1.14</td>
<td>.8</td>
<td>.95</td>
<td>.49</td>
</tr>
<tr>
<td>60</td>
<td>120</td>
<td>.5</td>
<td>.50</td>
<td>.00</td>
<td>.21</td>
<td>1.01</td>
<td>.1</td>
<td>1.01</td>
<td>.1</td>
<td>.97</td>
<td>.50</td>
</tr>
<tr>
<td>60</td>
<td>120</td>
<td>.5</td>
<td>.50</td>
<td>.00</td>
<td>.21</td>
<td>.79</td>
<td>2.5</td>
<td>.69</td>
<td>2.0</td>
<td>1.47</td>
<td>.64</td>
</tr>
<tr>
<td>60</td>
<td>120</td>
<td>.5</td>
<td>.50</td>
<td>.00</td>
<td>.21</td>
<td>.95</td>
<td>.5</td>
<td>.89</td>
<td>-.6</td>
<td>1.12</td>
<td>.54</td>
</tr>
<tr>
<td>60.0</td>
<td>120.0</td>
<td>.5</td>
<td>.50</td>
<td>.00</td>
<td>.21</td>
<td>.95</td>
<td>.5</td>
<td>.89</td>
<td>-.6</td>
<td>1.12</td>
<td>.54</td>
</tr>
<tr>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>1.3</td>
<td>1.5</td>
<td>1.8</td>
<td>1.2</td>
<td>.08</td>
<td>.08</td>
</tr>
<tr>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>1.5</td>
<td>1.6</td>
<td>2.0</td>
<td>1.3</td>
<td>.09</td>
<td>.09</td>
</tr>
</tbody>
</table>

Model, Populn: RMSE .21 Adj (True) S.D. .00 Separation .00 Reliability 1.00
Model, Sample: RMSE .21 Adj (True) S.D. .00 Separation .00 Reliability .80
Model, Fixed (all same) chi-square: .0 d.f.: 4 significance (probability): 1.00

**Appendix A: FACETS output**

**i) Pseudo candidate scripts’ output**

---

**RESEARCH MATTERS:**

8 | RESEARCH MATTERS : ISSUE 14 / JUNE 2012
### Table 7.1.1 Judge Measurement Report (arranged by mN)

<table>
<thead>
<tr>
<th>Score</th>
<th>Count</th>
<th>Average</th>
<th>Measure</th>
<th>S.E.</th>
<th>MnSq ZStd</th>
<th>MnSq ZStd</th>
<th>Discrm</th>
<th>PtMea</th>
<th>PtExp</th>
<th>Nu Script</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.5</td>
<td>25</td>
<td>.5</td>
<td>1.00</td>
<td>.65</td>
<td>.83</td>
<td>.85</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>12.5</td>
<td>25</td>
<td>.5</td>
<td>1.00</td>
<td>.75</td>
<td>.75</td>
<td>.85</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.97</td>
<td>1.26</td>
<td>1.1</td>
<td>.13</td>
<td>1.28</td>
<td>1.04</td>
<td>1.28</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.95</td>
<td>.87</td>
<td>.87</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.99</td>
<td>1.04</td>
<td>.91</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.99</td>
<td>1.01</td>
<td>.91</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.99</td>
<td>1.07</td>
<td>.91</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.99</td>
<td>1.07</td>
<td>.91</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.99</td>
<td>1.01</td>
<td>.91</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.99</td>
<td>1.01</td>
<td>.91</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.99</td>
<td>1.07</td>
<td>.91</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
</tbody>
</table>

### Table 7.1.2 Script Measurement Report (arranged by mN)

<table>
<thead>
<tr>
<th>Score</th>
<th>Count</th>
<th>Average</th>
<th>Measure</th>
<th>S.E.</th>
<th>MnSq ZStd</th>
<th>MnSq ZStd</th>
<th>Discrm</th>
<th>PtMea</th>
<th>PtExp</th>
<th>Nu Script</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.5</td>
<td>25</td>
<td>.5</td>
<td>1.00</td>
<td>.85</td>
<td>.83</td>
<td>.85</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>12.5</td>
<td>25</td>
<td>.5</td>
<td>1.00</td>
<td>.75</td>
<td>.75</td>
<td>.85</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.97</td>
<td>1.26</td>
<td>1.1</td>
<td>.13</td>
<td>1.28</td>
<td>1.04</td>
<td>1.28</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.95</td>
<td>.87</td>
<td>.87</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.99</td>
<td>1.04</td>
<td>.91</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.99</td>
<td>1.07</td>
<td>.91</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.99</td>
<td>1.07</td>
<td>.91</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.99</td>
<td>1.01</td>
<td>.91</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.99</td>
<td>1.01</td>
<td>.91</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.99</td>
<td>1.07</td>
<td>.91</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>.5</td>
<td>1.00</td>
<td>.99</td>
<td>1.07</td>
<td>.91</td>
<td>-1</td>
<td>.23</td>
<td>-3</td>
<td>1.19</td>
</tr>
</tbody>
</table>

### RESEARCH MATTERS: ISSUE 14 / JUNE 2012

ii) Real candidate scripts' output

Table 7.1.1 Judge Measurement Report (arranged by mN)

With extremes, Model, Populn: RMSE .24 Adj (True) S.D. .24 Separation .00 Reliability .90
With extremes, Model, Sample: RMSE .24 Adj (True) S.D. .24 Separation .00 Reliability .80
Model, Fixed (all same) chi-square: .0 d.f.: 4 significance (probability): .00

With extremes, Model, Populn: RMSE .57 Adj (True) S.D. .98 Separation .01 Reliability .98
With extremes, Model, Sample: RMSE .57 Adj (True) S.D. .98 Separation 7.01 Reliability .98
Without extremes, Model, Populn: RMSE .48 Adj (True) S.D. .98 Separation 8.29 Reliability .99
Without extremes, Model, Fixed (all same) chi-square: 2125.0 d.f.: 33 significance (probability): .00
With extremes, Model, Random (normal) chi-square: 32.5 d.f.: 33 significance (probability): .00

With extremes, Model, Populn: RMSE .57 Adj (True) S.D. .98 Separation .01 Reliability .98
With extremes, Model, Sample: RMSE .57 Adj (True) S.D. .98 Separation 7.01 Reliability .98
Without extremes, Model, Populn: RMSE .48 Adj (True) S.D. .98 Separation 8.29 Reliability .99
Without extremes, Model, Fixed (all same) chi-square: 2125.0 d.f.: 33 significance (probability): .00
With extremes, Model, Random (normal) chi-square: 32.5 d.f.: 33 significance (probability): .00

RESEARCH MATTERS: ISSUE 14 / JUNE 2012
The effect of scripts’ profiles upon comparability judgements

Nicky Rushton Research Division

Introduction

Examinations in England usually require candidates to take multiple units, with the results being combined to give their overall grade. This system allows for compensation. Using a mark scheme, a candidate’s answer to each question is scored individually and then aggregated to provide a total score for each component of the examination. A weak answer to one question may be compensated by a strong answer to another. Similarly, when several units are combined it is usual for the performances to vary across each one, sometimes by a considerable amount. Whilst this is accommodated in the marking and aggregation process, use of these candidate performances in activities requiring holistic judgement, such as comparability studies,1 can be problematic. Holistic judgement depends on the ‘whole picture’ being compared, and judges who are unduly influenced by the best answer or the worst answer may struggle. Equally, it is likely to be much harder to compare an uneven (or erratic) performance with one where the candidate performance entirely coheres. Comparing two performances which are uneven in different ways is an even more complicated task. This study aimed to investigate the effect of including candidates with uneven profiles in comparability studies. It was hypothesised that using uneven profiles may make the task of judging performance more difficult, and that it may affect judges’ perception of script quality.

Using scripts with even (or balanced) profiles for comparability studies is thought to make the judgements easier for examiners (Elliott and Greatorex, 2002; Pollitt and Elliott, 2003). Several judgemental studies have reported that examiners have found it more difficult to compare scripts with uneven profiles (Cresswell, 1997; Edwards and Adams, 2002; Scharaschkin and Baird, 2000). However, candidates with even profiles are uncommon (Elliott and Greatorex, 2002). Pseudo-candidates’ scripts (where scripts from more than one candidate are combined to create even profiles) may be generated if ‘true’ candidates are not available, but they may make the judgement task more difficult (Pollitt and Elliott, 2003; Yim and Forster, 2010), as a result of differences in writing style, tone and performance profile (Yim and Forster, 2010).

Whilst it is suggested that even profile scripts should be used for comparability studies, there are no widely accepted definitions of even and uneven profiles. Scharaschkin (1997) described an even profile candidate at the grade E boundary as one who had achieved “... (close to) the grade E boundary mark on each component” (p.1), although he also suggested that such candidates could also be defined using percentiles or z-scores. The percentiles method was used in a study where the profiles were defined statistically by calculating the range of marks achieved on the questions in the examination for every script (Scharaschkin and Baird 2000). Consistent (or even) scripts had a range of marks at the 5th percentile of the cohort or less, average scripts had a range of marks at the fiftieth percentile, and inconsistent (or uneven) scripts were at the ninety-fifth percentile. Bramley (2012) used the misfit statistic from Rasch analysis to identify high fitting candidates (those achieving the greater proportion of their marks on the easier questions) and low fitting candidates (those achieving the greater proportion of their marks on the harder questions). Crisp (2010) described unbalanced scripts (uneven profiles) as scripts where candidates had a higher score on one of the two essays than the other although the actual difference in marks was not described. Elliott and Greatorex (2002) suggested that an even profile was one where the scores on each component of an examination were balanced, that is the candidate performs equally well on each component of the examination. Edwards and Adams (2002) suggested that imbalanced (or uneven) scripts had missing questions, misread questions or rubric infringements.

Although there appears to be no single accepted definition of an even profile, a few studies have investigated the use of even profile candidates in judgemental tasks. This research has tended to focus on judgements about individual scripts rather than comparisons between pairs or groups of scripts. Scharaschkin and Baird (2000) found that uneven profiles affected examiners’ grading of candidates in A level biology and sociology with a significant effect at the grade A and grade E boundaries in biology, and at the grade A boundary in sociology. In all three cases, candidates with uneven profiles were less likely to be judged worthy of the higher grade. Crisp (2010) discussed uneven profiled candidates in a study investigating the features of candidates’ work that influenced grading decisions in A level geography. Several of the examiners reported that the A2 unit had been difficult to grade because candidates had performed better on one essay than the other. Contention emerged over whether to reward ‘spark’ or to base grading decisions on an impression of whether a grade A was deserved across the whole unit.

Only one study appears to have investigated the use of even profiles within a comparability study. Bramley (2012) investigated the effect of modifying four script features, including the profile of the script, within a rank ordering study. Candidates’ chemistry scripts were analysed using the Rasch model, and scripts with uneven profiles were selected. The answers which caused the uneven profile were identified, and answers that more closely matched the profile of the script substituted from other candidates’ scripts to create a more balanced script with the same total mark. Both the manipulated and the original script were then used in a rank ordering study. The results showed that scripts where a greater proportion of the marks were achieved on more difficult questions were perceived as better, but that this was affected by the proportion of the marks gained on questions considered to assess good chemistry. It was
suggested that the profile of scripts should be considered when choosing scripts for holistic judgement.

To date there is no evidence to show how uneven profiles affect judgements when scripts from more than one component of an examination are used. This study aimed to provide evidence in response to two research questions:

1. Are scripts with uneven profiles judged more harshly than those with even profiles in comparability studies?
2. Do judges give comparisons between even profiled scripts easier ratings for difficulty?

This study extended the Scharaskin and Baird (2000) results to syllabus level comparisons. The definitions of even and uneven profiles used within the study will be explored below.

**Method**

Most recent comparability studies have used rank ordering, where judges are presented with a selection of scripts to place in order, as it generates more information from fewer judgements. However, it was thought that the cognitive load placed upon judges would be too great for this particular task. Therefore this study used Thurstone’s paired comparison method (Thurstone, 1927; see also Bramley, 2007) where judges make judgements about the relative quality of pairs of scripts in order to compare examinations.

Two OCR A level specifications, Chemistry (H434) and English Literature (H471), were chosen for the study as they offered contrasting styles of assessment. (English Literature was assessed by essays, whilst Chemistry was assessed by shorter, more structured questions.) A level Chemistry consisted of six units, four of which were externally assessed. A level English Literature consisted of four units, two of which were externally assessed. Only the externally assessed units were used, because all the scripts for these units were available. In English Literature units F661 and F663 were chosen. In Chemistry the two A2 units, F324 and F325, were included in the study.

**Defining even and uneven profiles**

Differences in performance profiles may be observed in scripts as follows:

- Between the performances on units/components, e.g. candidate obtains a B overall comprising a unit at A and a unit at C.
- Between the performances on sub-components of a unit/component, e.g. candidate obtains a grade B overall on a unit/component, with a strong performance on the multiple-choice sub-component and a weak performance on the practical test sub-component.
- Between the performances on different sections of a unit. These may test different skills or knowledge, and candidates may be stronger in one area than another.
- Between the performances on different questions. Individual candidates’ performances may vary between different questions for a huge number of reasons.

This study was concerned with the first of these differences in performance profile. An even profile was defined as one in which the candidate had received the same grade in the units used within the study, fitting Elliott and Greatorex’s (2002) definition of an even profile.

An uneven profile was defined as one where a candidate had a range of two grades in their results (e.g., A, C). This definition of an uneven profile was chosen because it was not uncommon amongst candidates, and it was of interest having been used in some rank ordering studies.

**Script selection**

Scripts were selected from candidates who had taken both units in the June 2010 session, who fitted the profile criteria. Where possible, candidates were selected with a balanced performance within the unit to eliminate the profile within the unit as an extra factor which could influence the results.

Scripts on the same total mark but with different mark profiles were selected. Even profile candidates received the same grade on each unit (e.g., BB), whilst uneven profile candidates achieved their total mark in two possible ways (e.g., AC or CA). It was thought important to investigate both possible uneven profiles to see whether the perception of quality was affected by the unit within which the higher performance occurred. Two even profiled scripts were used so that there were equal numbers of even and uneven profiled scripts for each mark. These four scripts enabled six possible comparisons to be made for each total mark, as shown in Figure 1. Three script samples were selected for three total marks at grades B, C, and D, producing nine sets of four scripts.

![Figure 1: Comparisons of scripts at a single mark point](image1)

Comparisons were also carried out between scripts with different marks within the range for that grade, to see whether the total mark had an effect upon how scripts were judged. These additional comparisons were made between all the scripts at two of the different marks within a grade, (e.g. 93 and 95 marks were chosen in the grade B range) as shown by the solid lines in Figure 2.

![Figure 2: Comparisons of scripts with two different marks](image2)

2. In this article, ‘script’ refers to the whole candidate work being considered together, and comprises the answer to more than one unit.
Script preparation
The scripts were cleaned of any marks and annotations. They were then photocopied and given anonymous identification numbers which did not relate to the total mark that they received. The scripts were assembled into packs, such that the twelve scripts at a particular grade were assembled into the same pack. There were three packs for each subject (one for each grade B to D).

Judges
Ten judges were recruited for the study, five from each subject. All were senior examiners for that specification. It was not known whether the judges had completed paired comparisons before, but some of the judges had recent awarding experience, which would have required them to make holistic judgements of script quality.

Task
All the judges within a subject received identical packs. This enabled the consistency of examiners’ decisions to be investigated, as it had been noted in previous research that examiners did not treat uneven scripts consistently. A recording sheet was provided listing all the possible comparisons within a pack. They were listed in a different order for each judge to avoid order effects and the order was specified to ensure that no script was retained for more than two consecutive comparisons.

The instructions asked the judges to decide which script in a pair represented the better performance and ring that script on the recording sheet. They were then asked to decide how easy it was to make the judgement using a scale of 1 (very easy) to 5 (very difficult). This process was repeated for each of the script pairs within every pack.

Questionnaires
The judges were also sent questionnaires probing different aspects of the comparison process, such as how they made their judgements and what made a comparison difficult. No direct questions were asked about the impact of the profile of scripts but it was expected that if the profile of the script was an issue during comparisons, this would be raised within the responses to the questions about how judgements were made and what made comparisons difficult.

Results
Chemistry: comparison of profiles between pairs of units
When an even profiled script was compared to an uneven profiled script (Table 1), the even profiled scripts were slightly more likely to lose their comparisons. This was true for the results of four of the judges, although one judge’s results showed a tendency for even profiled scripts to win more comparisons. The significance of the even profiled scripts losing was explored using a binomial test. There was no significant effect of the script profile for all the judgements combined (p>0.05).

Table 1: Overall judgements involving even profile scripts

<table>
<thead>
<tr>
<th>Judge 1</th>
<th>Judge 2</th>
<th>Judge 3</th>
<th>Judge 4</th>
<th>Judge 5</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Even profile wins</td>
<td>27</td>
<td>29</td>
<td>33</td>
<td>27</td>
<td>26</td>
</tr>
<tr>
<td>Even profile loses</td>
<td>33</td>
<td>31</td>
<td>27</td>
<td>33</td>
<td>34</td>
</tr>
</tbody>
</table>

There was very little difference between even and uneven scripts winning comparisons when the results for scripts with the same total scores were compared (Table 2). Judge 3 seemed to slightly favour even profile scripts whilst Judge 4 seemed to slightly favour uneven profile scripts (20 wins). The binomial test for significance showed that there was no significant effect of the script profile for all the judgements combined (p>0.05).

Table 2: Comparison of scripts with the same total scores

<table>
<thead>
<tr>
<th>Judge 1</th>
<th>Judge 2</th>
<th>Judge 3</th>
<th>Judge 4</th>
<th>Judge 5</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Even profile wins</td>
<td>17</td>
<td>17</td>
<td>21</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>Even profile loses</td>
<td>19</td>
<td>19</td>
<td>15</td>
<td>20</td>
<td>19</td>
</tr>
</tbody>
</table>

Interestingly, for the comparisons where scripts had different scores, the scripts with higher scores lost more comparisons than they won (Table 3), both for the overall judgements and also for all the individual judges. This result was not expected, and suggested that the judges’ decisions may have been based on a feature of the script packs other than overall performance. The binomial test for significance showed that the overall difference in the number of times that the even higher total score lost was significant (p<0.01).

Table 3: Comparison by total score

<table>
<thead>
<tr>
<th>Judge 1</th>
<th>Judge 2</th>
<th>Judge 3</th>
<th>Judge 4</th>
<th>Judge 5</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higher score wins</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Lower score wins</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

To investigate whether this difference arose from the profile of the script, the comparisons by score were broken down according to whether the winning script had an even profile (Table 4). For most judges, the results were fairly evenly distributed, with roughly equal proportions of even and uneven profile scripts winning the comparisons, both when the higher score won and when the lower score won. These results suggested that the profile of the scripts was not responsible for the higher scoring scripts losing the majority of their comparisons, as the differences between the results from even and uneven profiled scripts were relatively small. Judge 5’s results differed from the others as they contained a slightly higher proportion of uneven profile scripts winning, regardless of whether its score was higher. The binomial test for significance showed that there was no significant effect of the script profile when broken down by score for all the judgements combined (p>0.05).

Table 4: Comparison of scripts by profile and total score

<table>
<thead>
<tr>
<th>Winning script</th>
<th>Judge 1</th>
<th>Judge 2</th>
<th>Judge 3</th>
<th>Judge 4</th>
<th>Judge 5</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higher score wins</td>
<td>Even profile wins with higher score</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Uneven profile wins with higher score</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>24</td>
</tr>
<tr>
<td>Lower score wins</td>
<td>Even profile wins with lower score</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Uneven profile wins with lower score</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>42</td>
</tr>
</tbody>
</table>
Finally, the Chemistry results were analysed according to the profiles of the scripts used in the comparison (Table 5). Three profiles were used in the comparisons: an even profile where the same grade was achieved on both units; an uneven profile where the candidate’s result was two grades higher on the F324 script than it was on the F325 script (better F324 result); an uneven profile where candidate’s grade was two grades higher on the F325 script than it was on the F324 script (better F325 result).

There were some variations in the judges’ decisions. When a better F324 result was compared with an even profile candidate, the results from Judges 1 and 3 showed that the even profile scripts won and lost an equal number of times, whereas Judge 4 and 5’s results showed that the uneven profile scripts won slightly more often, but again there was not much difference between the two figures. The binomial test for significance showed that there was no significant effect of the script profile for all the judgements combined (p>0.05).

For the comparisons where a better F325 result was compared with an even profile script, the judges’ decisions varied slightly more. The results from Judges 1, 2 and 5 seemed to favour the better F325 result scripts over the even profile script, whereas Judge 3’s decisions seemed to favour the even profile scripts. The binomial test for significance showed that there was no significant effect of the script profile for all the judgements combined (p>0.05).

In both sets of comparisons for the uneven vs. even profiles, the majority of the decisions suggested that the uneven profile scripts were judged slightly more favourably than the even profiled scripts, although the decisions made by individual judges did not necessarily follow the same pattern in both sets of judgements. For example, Judge 2 seemed to slightly favour even profile scripts over the better F324 scripts, but then favoured the better F325 scripts over the even profile scripts. When the two uneven profiles were compared to each other there did not seem to be much difference in the number of times each type of profile won its comparisons, suggesting that there was not really any difference between the two types of uneven profile when it came to forming judgements. The binomial test for significance showed that there was no significant effect of the script profile for all the judgements combined (p>0.05).

<table>
<thead>
<tr>
<th>Table 5: Comparison by profile of script</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge 1</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Even vs. better F324 result</td>
</tr>
<tr>
<td>Even profile wins</td>
</tr>
<tr>
<td>Better F324 wins</td>
</tr>
<tr>
<td>Even vs. better F325 result</td>
</tr>
<tr>
<td>Even profile wins</td>
</tr>
<tr>
<td>Better F325 wins</td>
</tr>
<tr>
<td>Better F324 vs. better F325 result</td>
</tr>
<tr>
<td>Better F324 wins</td>
</tr>
<tr>
<td>Better F325 wins</td>
</tr>
</tbody>
</table>

English Literature: comparison of profiles between pairs of units

In the English Literature comparison, when even profiled scripts were compared with uneven profiled scripts, the even profile scripts lost more comparisons than they won (Table 6). Only Judge E’s results deviated from this profile, as even profile scripts won and lost roughly equal numbers of comparisons. The significance of the even profiled scripts losing was explored using a binomial test. There was a significant effect of the script profile for all the judgements combined (p<0.01).

Table 6: Overall judgements involving even profile scripts

<table>
<thead>
<tr>
<th>Judge 1</th>
<th>Judge 2</th>
<th>Judge 3</th>
<th>Judge 4</th>
<th>Judge 5</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Even profile wins</td>
<td>17</td>
<td>24</td>
<td>24</td>
<td>21</td>
<td>28</td>
</tr>
<tr>
<td>Even profile loses</td>
<td>39</td>
<td>32</td>
<td>32</td>
<td>35</td>
<td>27</td>
</tr>
</tbody>
</table>

When the scores were the same (Table 7), most of the judges’ results showed a considerable bias towards the uneven profiled scripts, although Judge E treated both the same. The significance of the even profiled scripts losing was explored using a binomial test. The binomial test for significance showed that there was a significant effect of the script profile for all the judgements combined (p<0.01).

Table 7: Comparison of scripts with the same total scores

<table>
<thead>
<tr>
<th>Judge A</th>
<th>Judge B</th>
<th>Judge C</th>
<th>Judge D</th>
<th>Judge E</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Even profile wins</td>
<td>10</td>
<td>13</td>
<td>14</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>Even profile loses</td>
<td>24</td>
<td>21</td>
<td>20</td>
<td>20</td>
<td>17</td>
</tr>
</tbody>
</table>

The influence of the total score across both units used for the judgements was also investigated (Table 8). In this analysis, the combined results from all the judges showed that neither score was favoured because the results from the individual judges cancelled each other out. Judges A, B and C’s results slightly favoured the higher scoring scripts, whilst Judges D and E’s results slightly favoured the lower scoring scripts. No statistical significance test was carried out on these results, as no difference between the categories was observed for all the judges combined.

Table 8: Comparison by total score

<table>
<thead>
<tr>
<th>Judge A</th>
<th>Judge B</th>
<th>Judge C</th>
<th>Judge D</th>
<th>Judge E</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higher score wins</td>
<td>13</td>
<td>13</td>
<td>12</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Lower score wins</td>
<td>9</td>
<td>9</td>
<td>10</td>
<td>13</td>
<td>14</td>
</tr>
</tbody>
</table>

The results by score were then broken down to see whether there were any patterns in the profile of the scripts that might help to explain why some judges favoured the lower scoring scripts (Table 9). For Judges B, C and E the results were fairly evenly distributed, with roughly equal proportions of even and uneven profiled scripts winning the comparisons, both when the higher score won and when the lower score won. This suggested that for the majority of the judges the profile of the scripts did not affect whether the higher scoring script won and the binomial test for significance confirmed that there was no significant effect of the script profile when broken down by score for all the judgements combined (p>0.05). However, Judges A and D results showed a higher proportion of uneven profile scripts winning their comparisons, regardless

3. There are only 16 comparisons per judge in English Literature study. This is because one of the scripts had to be removed from the comparisons as it consisted of two copies of unit F661, rather than F661 and F662.
of whether the uneven profile script had a higher score. Therefore, for these two judges the profile of the script may have mattered slightly more than whether the script’s score was higher.

Table 9: Comparison of scripts by profile and total score

<table>
<thead>
<tr>
<th></th>
<th>Judge A</th>
<th>Judge B</th>
<th>Judge C</th>
<th>Judge D</th>
<th>Judge E</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Higher score wins</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Even profile wins</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>Uneven profile wins</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>34</td>
</tr>
<tr>
<td><strong>Lower score wins</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Even profile wins</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>26</td>
</tr>
<tr>
<td>Uneven profile wins</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>6</td>
<td>29</td>
</tr>
</tbody>
</table>

Finally, the results were analysed to see whether the unit that had a better performance made any difference to the results (Table 10). Three profiles were used: an even profile where the same grade was achieved on both units; an uneven profile where the candidate achieved two grades more on the A2 unit than they did on the AS unit (better A2 result); and an uneven profile where candidates achieved two grades more on the A2 unit than they did on the AS unit (better A2 result).

There were some variations in the decisions that the judges made about the scripts. When a better AS result script was compared with an even profile script, all the judges decided that the uneven profile script showed the better performance more frequently than the even profile script. For most judges the difference between the number of times the even profile script won and lost the comparisons was a reasonably large one; only Judge E was close to judging even profile scripts winning and losing an equal number of times. The binomial test for significance showed that there was a significant effect of the script profile for all the judgements combined (p<0.01).

Table 10: Comparison by profile of script

<table>
<thead>
<tr>
<th></th>
<th>Judge A</th>
<th>Judge B</th>
<th>Judge C</th>
<th>Judge D</th>
<th>Judge E</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Even vs. better AS result</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Even profile wins</td>
<td>7</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>14</td>
<td>48</td>
</tr>
<tr>
<td>Even profile loses</td>
<td>19</td>
<td>15</td>
<td>17</td>
<td>19</td>
<td>12</td>
<td>82</td>
</tr>
<tr>
<td><strong>Even vs. better A2 result</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Even profile wins</td>
<td>10</td>
<td>13</td>
<td>15</td>
<td>14</td>
<td>14</td>
<td>66</td>
</tr>
<tr>
<td>Even profile loses</td>
<td>20</td>
<td>17</td>
<td>15</td>
<td>16</td>
<td>15</td>
<td>83</td>
</tr>
<tr>
<td><strong>Better A5 result vs. better A2 result</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Better A5 result wins</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>11</td>
<td>7</td>
<td>40</td>
</tr>
<tr>
<td>Better A2 result wins</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>24</td>
</tr>
</tbody>
</table>

For the candidates with a better A2 result, the judges’ decisions were more inconclusive. Judges A and B appeared to favour the uneven profile scripts (those with the better A2) result over the even profile scripts. Judges C, D and E appeared not to favour either profile. The binomial test for significance showed that there was no significant effect of the script profile for all the judgements combined (p>0.05).

The results of the comparisons between scripts with a better AS performance and those with a better A2 performance showed that the judges tended to judge both scripts as winning about the same number of times. The only exception was Judge D who showed a strong tendency to favour the scripts with a better performance on the AS unit. The binomial test for significance showed that there was a significant effect of the script profile for all the judgements combined (p<0.05).

### Perceived difficulty of making Chemistry judgements

Information about the difficulty of making judgements was taken from two sources: the judges’ ratings for the difficulty of making each paired comparison and the judges’ responses to the questionnaires. All of the Chemistry judges rated their paired comparisons using the whole range of the 5 point rating scale from 1 (easy) to 5 (very difficult). Two of the comparisons were not given difficulty ratings, but it was thought that this was an oversight as the two occurrences came from different judges.

An average was taken of each Chemistry judge’s ratings for each type of profile involved in the comparison, and then these were totalled across particular types of comparison (Table 11).

Table 11: Difficulty ratings of Chemistry judges

<table>
<thead>
<tr>
<th></th>
<th>Judge 1</th>
<th>Judge 2</th>
<th>Judge 3</th>
<th>Judge 4</th>
<th>Judge 5</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Even profile vs. better F324 result</td>
<td>3.1</td>
<td>3.2</td>
<td>3.7</td>
<td>3.3</td>
<td>2.9</td>
<td>3.3</td>
</tr>
<tr>
<td>Even profile vs. better F325 result</td>
<td>3.0</td>
<td>3.3</td>
<td>3.3</td>
<td>3.3</td>
<td>2.7</td>
<td>3.1</td>
</tr>
<tr>
<td>Even profile vs. even profile</td>
<td>2.6</td>
<td>3.0</td>
<td>3.4</td>
<td>2.8</td>
<td>2.1</td>
<td>2.8</td>
</tr>
<tr>
<td>Better F324 result vs. better F325 result</td>
<td>3.7</td>
<td>3.4</td>
<td>3.3</td>
<td>3.1</td>
<td>3.6</td>
<td>3.4</td>
</tr>
<tr>
<td>Both scripts better F324 result</td>
<td>5.0</td>
<td>3.7</td>
<td>4.0</td>
<td>4.7</td>
<td>3.0</td>
<td>4.1</td>
</tr>
<tr>
<td>Both scripts better F325 result</td>
<td>4.3</td>
<td>3.7</td>
<td>2.7</td>
<td>1.7</td>
<td>0.7</td>
<td>2.6</td>
</tr>
</tbody>
</table>

The Chemistry judges tended to rate all of the types of comparison towards the upper end of the difficulty scale. Those judgements involving even profiled scripts did not stand out as more difficult or easier than the other comparisons in either the combined ratings, or in the ratings given by individual judges.

The judges’ questionnaire responses were also analysed to find out what they thought made the comparisons difficult. All of the judges commented that the profile of the scripts influenced their judgements. These comments fell into several categories:

- Comments about different performance across the two units which occurred in the feedback from every single judge, suggesting that whilst comparisons involving uneven profile scripts did not appear to have received higher difficulty ratings, they were perceived to be more difficult.

  **Different performance across F324 and F325** [Judge 4]

- Comments concerning inconsistency of performance within a unit. Some of the judges referred to this type of inconsistency within their answers about what made comparisons difficult.

  **Candidates who showed inconsistency, i.e. some very good answers followed by inadequate and seemingly ignorant responses**. [Judge 1]
• Comments referring to missing answers or the candidate failing to finish the unit, which may have contributed to the perception of an unbalanced script.

Scripts with significant gaps that made the comparison between a smaller number of ‘better answers’ with a larger number of ‘poorer answers’ [Judge 4]

I also looked for instances where a candidate gave an answer which was clearly unfinished or showed a lack of understanding [Judge 1]

Some of these causes of uneven performances would not have been identified during the script selection process.

The judges’ comments on how they made their decisions also revealed that they all focused on one unit more than another, although their reasons for doing so varied.

I regard F325 as a more realistic indicator of real ability in Chemistry [Judge 1]

F325 seemed to be a much better indicator [Judge 2]

They all also indicated that they referred to some questions more than others when they were making their decisions. There were a variety of reasons for choosing particular questions.

I identified the questions on each paper that I felt demonstrated good understanding and placed more emphasis on these rather than those which required direct recall. [Judge 1]

Questions that included stretch and challenge were considered less important [Judge 4]

Perceived difficulty of making English Literature judgements

All of the judges rated the majority of their paired comparisons using the whole range of the 5 point rating scale from 1 (easy) to 5 (very difficult). There were a few comparisons where ratings had not been given, but there were few enough instances of this for analysis of the overall results to be possible. An average was taken of each judge’s ratings for every combination of profiles involved in the comparison. These were then totalled across particular types of comparison.

The difficulty ratings given by the English Literature judges were fairly similar for all the types of comparisons (Table 12). The comparisons between even profiled scripts and scripts with the better performance for the AS unit were given a marginally more difficult rating overall, but this difference was too small to indicate that there was any real difference in difficulty. No judge rated this as the most difficult type of comparison, nor was there any agreement amongst their responses to suggest what type of comparison was most difficult. Two judges’ scores suggested that the comparisons between scripts where both scripts had an even profile were the most difficult. Two further judges’ scores suggested that the comparisons where both scripts had a better A2 result were the most difficult, and the remaining judge’s scores suggested that comparisons between even profile scripts and scripts with a better A2 result were the most difficult. There was nothing to indicate that comparisons involving even profiled scripts were either more difficult or easier than other comparisons.

The judges’ questionnaire responses were also analysed to see what they thought made the comparisons difficult. Like the Chemistry judges, several of the questionnaire responses contained comments on the profile of the scripts. However, the issues raised by the English Literature judges seemed subtly different. Several of the judges commented on the uneven profile of the scripts in response to the question about what made some comparisons difficult.

Weaker scripts are more difficult to judge because they can have brief flashes of coherence [Judge C]

Parts of scripts in which the same candidate had performed very differently in each unit [Judge D]

They also tended to comment on a perceived closeness in the quality of the scripts in their responses to the difficulty question.

Similarity of performance… the closer the pairs in performance, the harder the decision [Judge A]

In this set of English Literature comparisons, one of the judges’ questionnaire responses indicated that features of the scripts other than the intended grade differences between units may have led to a perception of uneven performance.

…[Holistic judgement] depends upon no unevenness in the performance, the mark profile or the weight of copy across the script [Judge E]

This judge was the only one to draw attention to inconsistency in features such as the amount of writing.

Three of the English Literature judges considered both units equally, but two judges commented that they did not do so.

… I tended to make a preliminary judgement based on F661: this is the unit with which I am more familiar [Judge D – English Literature]

One of the judges also indicated that they considered some questions more than others.

‘Both units were considered equally, but section A in F661… and section A in F663 were very much more important” [Judge A – English Literature]
Discussion

Analysis showed that the effect of including uneven profile candidates depended on both the subject and the judges. An effect of profile was seen in some of the comparisons, but the effect of the script profile appeared to be less consistent in Chemistry than it was in English Literature. Tests of statistical significance were carried out where the comparisons were analysed by the profile of the script (even or uneven). The effect of the profile was found to be statistically significant in English Literature, but not in Chemistry. The results were not analysed by individual judges, as that may have caused problems with multiple testing.

The combined results from all the judges in the Chemistry comparisons showed that even profile scripts were slightly less likely to win their judgements. However, there seemed to be little obvious difference in the combined results for all judges when scripts with the same scores were compared. The majority of the Chemistry judges’ results indicated that the profile of the script was not influencing their judgements. Two of the judges did show some differences, with one seeming to favour the even profiled scripts and the other the uneven profiled scripts. These differences were seen both within the overall comparisons and the comparisons between scripts with the same score.

In the English Literature comparisons the even profile scripts tended to be judged as being of poorer quality than the uneven profile scripts. This happened when the scripts being compared had the same score and also when the even profiled scripts had a higher score than the uneven profiled scripts. These results were seen in the individual results from four of the judges, where the uneven scripts lost more comparisons both in the overall comparisons and in the comparisons where both scripts had the same score.

There are several possible explanations for why there were more noticeable differences within the English Literature comparisons than there were within the Chemistry judgements. First, an analysis of the results from the individual judges showed that the Chemistry judges varied more in their results than did the English Literature judges (although both subjects had some judges who favoured even profiled scripts and others who favoured uneven profiled scripts). Possibly a different selection of judges would have produced different results, and the difference that was observed between the subjects was merely the result of the judges that were used for the study.

A second explanation was that the style of examination led to differences in the comparisons. The English Literature units consisted of two essay questions per unit whereas the Chemistry units consisted of several questions, each with multiple sub-questions. It is likely that the differences in performance for the uneven profile scripts would be more obvious when four questions were compared, as happened in the English Literature comparisons, than it would be when many more questions were involved, and fluctuations in response are less extreme (because each question carries fewer marks) and less noticeable. However, as all of the Chemistry judges commented on the difficulty of comparisons involving uneven profiles within their questionnaire responses this explanation is less likely to be the only cause of the differences that were observed between the subjects.

Another explanation could have been that the judges did not properly consider all of the answers for both of the units included in a pair; thus, their perception of whether the script had an even or uneven profile may not have been correct. There was some evidence of this in the questionnaire responses from the Chemistry judges who all reported focussing on the F325 scripts, giving a range of reasons for doing so. Only one of the English Literature judges expressly focused more on one unit, citing increased familiarity with it as the reason for doing so. Additionally, some of the judges focused on particular questions within units that they believed discriminated well between the scripts. If the judgements were based on evidence from a small section of the script, it is possible that the judges’ perceptions of which scripts contained an uneven profile may not have matched the scripts identified as such within the study. This may have led to the smaller effect of the profile within Chemistry.

A surprising finding from the Chemistry comparisons was that the lower scoring script won more comparisons than the higher scoring script. This finding was consistent across all judges. Whilst some of the English Literature judges’ results showed that the lower scoring scripts won more comparisons, the difference was not as great, and the judges’ results cancelled each other out. There is no obvious explanation for the surprising Chemistry result. It is possible that the judges were focusing on particular questions, and that performance on these did not reflect the overall performance. Alternatively, the Chemistry judges may have formed their judgements on the basis of particular skills or areas of knowledge that did not receive as many marks as other areas that the judges considered to be less important.

This study also investigated whether it was perceived to be easier to make comparisons when the scripts had even profiles. In Chemistry, the ratings suggested that the script’s profile (uneven or even) did not affect the perceived difficulty of the comparisons. There was no obvious pattern within the English Literature results, which suggested that in English Literature the difficulty of making comparisons varied according to the judge used, rather than just being a result of the profile. Neither the Chemistry nor the English Literature results from the difficulty ratings matched the questionnaire data, where all the judges had commented upon either even or uneven profile scripts affecting the difficulty of making judgements. The difference between judges’ perceptions of difficulty and the ratings that they gave the comparisons when they involved uneven profiles is interesting. The questionnaire data match the findings in the literature (e.g. Cresswell, 1997; Edwards and Adams, 2002) that it is perceived to be more difficult to make comparisons with uneven profiled scripts. However, the data about the difficulty of making each judgement contradict this. One possible explanation for this was that if the judges focused on particular units or particular questions within the unit they may have formed different impressions of the scripts as having even or uneven profiles to those intended. Alternatively, the judges may have defined uneven or unbalanced scripts in a different way. There was some evidence for this explanation in the Chemistry questionnaire responses. Many of the Chemistry judges mentioned different performance on the two units, but some also commented on parts of questions being missed out or candidates failing to finish. Both of these could have produced a perception of the script having an uneven profile. Neither of these criteria was used to select scripts or identified as a feature to control, so it is possible that some of the scripts that were identified in the study as having an even profile may have been identified as uneven if the extra criteria had been included. There was only limited evidence of a different definition of an even profile within the English Literature questionnaire responses, where one judge mentioned the amount of writing produced within answers as a cause of an uneven profile. As these Chemistry and English Literature judges possibly perceived uneven profiles in a different way, they may not have
recognised the uneven profile scripts identified within the study as being so. That would have affected their difficulty ratings, and may help to explain why comparisons involving uneven profile scripts were perceived as difficult yet did not receive high ratings for difficulty.

There are several limitations to this study. First, it was not possible to find out the examiners’ definitions of an uneven profile script. Had this been investigated it may have been possible to explain the difference between the questionnaire findings about the difficulty of making comparisons with uneven profiled scripts and the difficulty ratings. Another limitation is that the judges were not experienced in considering multiple units when making judgements. This may have made the comparisons difficult for them and meant that they did not have a consistent perception of what a better performance consisted of. Finally, the English Literature comparison involved one AS level unit and one A2 level unit. The different standards of the two scripts may have complicated the process of forming a holistic judgement of the quality of the scripts.

Conclusion

This study investigated the effect of including uneven profile scripts in comparability studies to see whether it made any difference to the judgements judges made.

It was found that the uneven profile scripts were slightly more likely to win their comparisons, but that this depended on the judge involved. Some judges appeared to be influenced by the higher standard of performance that was observed on part of an uneven profile script, and thus decided that it should win comparisons. Other judges were less decisive, or favoured the candidates who could sustain a balanced performance throughout a script or the scripts from multiple units.

There was mixed evidence on the perceived difficulty of making comparisons when scripts had uneven profiles. Whilst the judges generally identified uneven profiles as a source of difficulty in their questionnaire responses, the results from the difficulty of making the comparisons seemed to contradict this. There did not appear to be any evidence that the profile of the script affected the difficulty of making the comparisons in Chemistry, and the effect seemed to vary according to the judge used in English Literature.

An additional interesting finding of the study was that some of the judges appeared to have different views of what an uneven performance was. Some of these views were more complex than the definitions used within the study of an uneven profile as different grades across units, or different marks for questions within one unit. The judges mentioned additional features as sources of an uneven profile such as: incomplete answers; a mismatch between the language used and the concepts expressed within the answers; and differing lengths of answers. A few of these features had been mentioned in the Edwards and Adams (2002) comparability study as causes of difficulty when making comparisons, but were not identified as a focus for this study. Most of these additional features would be difficult to identify when selecting scripts for a comparability study, as it would be too time consuming to identify them.

If the profile of a script affects how that script is judged, then the outcome of comparability studies could be affected by the inclusion of uneven profile scripts. For example, if even profile scripts are seen as weaker this would suggest that evidence of ‘spark’ is unduly affecting judgement.

It is of concern that some of the judges’ decisions may have been based on features of the script packs other than overall performance. This may indicate that judges are not completing the holistic task in the intended way, which has implications for other contexts where holistic judgement is used.

References


Monitoring the difficulty of tiered GCSE components using threshold marks for grade C

Vikas Dhawan  Research Division

Introduction

GCSE results are reported on a grade scale from A* (highest) to G. Since this scale covers a wide range of attainment, many GCSEs are divided into two tiers, Foundation and Higher. The Foundation tier covers grades G to C and the more difficult Higher tier covers grades D to A*, with grade E often allowed for those candidates who just miss grade D. Centres enter candidates to either the Foundation tier or Higher tier, depending on the centres’ judgement of which level is most appropriate for their candidates.

Table 1 shows the grades available on the Foundation and Higher components of a tiered GCSE unit. Example raw grade thresholds – the minimum raw mark required for each grade – are also shown for one session’s examinations (both examinations were marked out of 60). The thresholds reflect the difficulty of the examinations. For example, looking at the grade C thresholds for the two example examinations, candidates had to score at least 33 marks on the Foundation examination for a grade C, but only needed 19 marks on the harder Higher examination. The overlapping grades C and D must represent the same standard of attainment on both the tiers so that the resulting award is fair to all candidates. Usually, some items (question sub-parts) are common to both the Foundation and Higher examinations, and candidates’ performance on these common items can help inform the grade-thresholds setting process.

Table 1: Grades available in GCSE tiered components and sample grade thresholds

<table>
<thead>
<tr>
<th>Higher tier grade</th>
<th>A*</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>Ungraded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higher tier – minimum raw mark for grade</td>
<td>46</td>
<td>38</td>
<td>28</td>
<td>19</td>
<td>15</td>
<td>13</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Foundation tier grade</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>Ungraded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foundation tier – minimum raw mark for grade</td>
<td>33</td>
<td>27</td>
<td>22</td>
<td>17</td>
<td>12</td>
<td>0</td>
</tr>
</tbody>
</table>

The process of setting grade thresholds is known as awarding. Awarding is carried out for each examination (i.e. each session) under the procedure laid down by the GCSE regulator in the Code of Practice (Ofqual, 2011), which states that the purpose of awarding is “to ensure that standards are maintained in each subject examined from year to year”. Thresholds are set by expert judgement informed by a wide range of statistical and candidate-performance evidence.

The study reported in this article aimed to explore simple ways of monitoring the relative difficulty of tiered components by considering the difference between the grade C thresholds. If the difficulty of the question papers is as intended, and the grade thresholds have been set correctly, the C threshold mark on the Foundation paper will be higher – as a proportion of the paper total – than the C-threshold mark on the Higher paper. This is because the Higher tier paper will contain a greater proportion of difficult items on which grade C candidates would not be expected to accrue many marks. In fact OCR, Cambridge Assessment’s GCSE awarding body, has set itself the demanding target of constructing Foundation and Higher tier question papers such that the C-threshold is at around 85% of the Foundation paper total, and around 40% of the Higher paper total, leading to a target difference between Foundation and Higher tier C-thresholds of 45 percentage points.

There are two basic reasons why the difference between C-thresholds might not be as intended: either one or both question papers might not have been at the target difficulty; or one or both of the C-thresholds might not have been set on the right mark. Table 2 summarises these two reasons in the context of a difference in grade C thresholds having been found to be smaller than designed, or even the wrong way round. Note that the existence of a large difference between the C-thresholds is not a foolproof indicator that all is well. Incorrectly targeted question papers, or incorrectly set grade thresholds, would be just as likely to increase the difference between the Foundation and Higher C-thresholds as to decrease it. Additionally, factors might combine to reduce or increase the difference in thresholds. However, the difference between the C-thresholds is easy to calculate and, when combined with other indicators, might prove useful for routine monitoring of the technical qualities of assessments. Thus in the present study we calculated these differences for two examination sessions, and we present our findings in this article.

Table 2: Potential reasons for a small – or even reverse – difference between the grade C thresholds of tiered examinations

<table>
<thead>
<tr>
<th></th>
<th>Foundation tier</th>
<th>Higher tier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test construction</td>
<td>Too difficult at C</td>
<td>Correct difficulty</td>
</tr>
<tr>
<td>Awarding</td>
<td>Low C-threshold to compensate</td>
<td>Correct threshold</td>
</tr>
<tr>
<td>Test construction</td>
<td>Correct difficulty</td>
<td>Too easy at C</td>
</tr>
<tr>
<td>Awarding</td>
<td>Correct threshold</td>
<td>High threshold to compensate</td>
</tr>
<tr>
<td>Test construction</td>
<td>Correct difficulty</td>
<td>Correct difficulty</td>
</tr>
<tr>
<td>Awarding</td>
<td>Threshold set too low</td>
<td>Correct threshold</td>
</tr>
<tr>
<td>Test construction</td>
<td>Correct difficulty</td>
<td>Threshold set too high</td>
</tr>
</tbody>
</table>

For a more theoretical discussion on the concepts related to tiered papers see Wheadon and Béguin (2010) and Good and Cresswell (1988a, 1988b, 1988c).
2. Method

The assessments selected for this study were from the GCSE tiered examinations conducted by the awarding body OCR in two sessions, June 2009 and June 2010. For each pair of Foundation and Higher tier components, the C threshold marks were expressed as percentages of the total mark available and the difference between these percentages was calculated (Foundation C-threshold minus Higher C-threshold). Table 3 shows an example of this calculation using one component-pair from the June 2009 session.

Table 3: Example of the calculation of the difference in grade C thresholds between a Foundation (F) and Higher (H) tier examination

<table>
<thead>
<tr>
<th>C threshold % of Paper total</th>
<th>Paper total</th>
<th>C threshold % of Paper total</th>
<th>Difference in % points</th>
</tr>
</thead>
<tbody>
<tr>
<td>F 33</td>
<td>H 19</td>
<td>F 60</td>
<td>H 60</td>
</tr>
</tbody>
</table>

3. Results

Figures 1 and 2 summarise the difference in percentage points between the Foundation and Higher tier grade C thresholds from the June 2009 (n=98) and June 2010 (n=115) sessions respectively. The tallest bar in the histogram in Figure 1 shows that, for the June 2009 component-pairs, the Foundation tier C-threshold was between 15 and 25 percentage points above the Higher tier C-threshold for just under 35% of the component pairs. A normal distribution curve is also shown on the graphs for reference. The vertical line at 0.0 on the x-axis shows the point where the C-thresholds were set at the same percentage of raw marks on both tiers; the vertical line at 45.0 percentage points on the x-axis shows OCR’s target difference.

The average difference between Foundation and Higher C-thresholds was approximately 20 percentage points in both sessions. Only a small – but apparently increasing – percentage of the component-pairs had C-thresholds which met OCR’s target of 45% percentage points between the C-thresholds. This is not in itself particularly alarming, since one purpose of setting targets is to encourage improvements, and there is no suggestion from these data that candidates received incorrect results.

The astute reader will have noticed that three components had ‘reversed’ C-thresholds, that is, the Foundation tier C-threshold was set at a lower percentage than the Higher tier C-threshold. These three component-pairs were fairly idiosyncratic and are now obsolete. However, further statistical information will now be presented for one of these component pairs as an example of a straightforward investigation that can be done to diagnose possible causes of a reverse – or small – difference in the C-thresholds. Straightforward investigations such as these can be carried out routinely for any component-pair where the threshold-difference is less than desired, and the results used to focus more thorough investigation and improvements.

Figure 3 shows example results of a simple statistical analysis of item facility values to investigate a component-pair found to have a smaller than desired difference between the Foundation and Higher C-thresholds, or even a reverse difference, as was the case here. The ‘facility’ of an item is the average mark scored on the item divided by the maximum for the item. For example, if candidates scored on average 1 mark on an item worth 2 marks, the facility of the item for these candidates is 0.5. Facilities must be between 0 (nobody scored anything) and 1 (everybody scored full marks). Some 10 items, worth a total of 16 marks, were on both the Foundation and Higher tier question papers. These common items are labelled C1, C2, etc., in Figure 3. The remaining items on the Foundation paper are labelled F1–F20, and the remaining items on the Higher paper are labelled H1–H19. Note that we do not expect the facility values of the common items to be the same on both tiers, since we expect the Higher tier to attract a more able entry than the Foundation tier – leading to higher facility values for the common items on the Higher tier.

Two key observations can be made from Figure 3. First, the high facility values for many of the common items show them to be amongst the easiest items on both papers; this is unexpected, since presumably the common items were intended to be pitched at the overlapping grades, and therefore to be amongst the hardest on the Foundation paper and the easiest on the Higher paper. Although the harder common items, C3, C8 and C10, did have lower facility values for Foundation candidates than Higher candidates, the majority of the common items had similar – high – facility values for both. Moreover, on the Foundation paper, the...
difference between the facility values of the easy common items and the non-common items was generally greater than the difference on the Higher paper – implying that the non-common Foundation items were generally harder than the non-common Higher items. This is the second key observation from Figure 3: many of the non-common Foundation items were apparently harder than the non-common Higher items. Thus in this example, it appears that issues relating to test construction were the major cause which led to the unexpected difference in the C thresholds. Further investigation – by the teams involved in setting these examinations – would be required to explain what caused the mis-targeting and improve future papers.

4. Discussion

In this article we have looked at a simple indicator for monitoring the difficulty-targeting of examinations in tiered GCSEs, namely the difference in percentage points between Foundation tier and Higher tier Grade C thresholds. It was argued that if the Foundation tier C-threshold was not at a considerably higher percentage mark than the Higher tier C-threshold, at least one of the examinations must have been mistargeted, or at least one of the thresholds must have been set on the wrong mark. We showed that a straightforward comparison of the facility values of items common to both examinations with the facility values of items on only one examination allowed us to diagnose the likely cause of an unexpected difference between C-thresholds. In the case that we examined, the likely main cause was that the common items were too easy and the non-common items on the Foundation examination were too hard. Further investigation, including a detailed review of the items, would be required to identify what caused this mis-targeting.

The reason for using grade C threshold comparisons and item facility values was that they are routinely available. The production of graphs such as Figure 1 was easily automated and so can be incorporated into routine quality monitoring, and item facility values are already available for initial investigation of any tiered examinations flagged by the C-threshold comparisons. More sophisticated statistical and psychometric techniques, such as Rasch analysis, are more accurate and powerful, and would more clearly indicate whether the C-thresholds were set at the same standard on both tiers, and the relative difficulty of the non-common items, but at the cost of increased complexity of production. Their use might best be reserved for investigating issues flagged by simpler means such as those used in this article. However, all statistical and psychometric techniques for comparing the difficulty of the examinations and the standard of the C-thresholds, whether simple such as facility values or more sophisticated such as Rasch analysis, depend on the common items for linking the examinations. This leads to several caveats. First, the common items may not fully represent the entire content of the examinations, and it is possible that candidates performed differently on the non-common items. Moreover, by trying to make the items suitable for both tiers, the item writers might have introduced features to the common items that caused candidates to perform differently on these items, or which prevented the items from discriminating well between candidates of different ability. Secondly, measurement error depends in part on the number of items in a test, and so the measurement error associated with candidates’ scores on the ‘sub-test’ comprised of the common items will likely be greater than the measurement error associated with candidates’ scores on the whole examination. Kolen and Brennan (2004) recommend that at least 20% of the items on two tests to be equated should be common. However, this is on the assumption that the examinee groups are not very different. In the
context of tiered examinations, we are dealing with different groups (we expect more able candidates to enter the higher tier) and, as Klein and Kolen (1985) (cited in Cook and Petersen, 1987) demonstrated, when examinee groups are different the proportion of items common to the tests become important. Thirdly, the common items are rarely in the same order on both examinations, and this might affect the difficulty of the items. For these reasons, the linking of the tiers via candidates’ marks on the common items should be treated with caution.

It should be noted that having many items with low facilities on a Foundation paper, or many items with high facility values on a Higher paper, does not necessarily mean that the papers were mistargeted: candidates might have entered for the wrong tier.

The comparison of grade C thresholds of tiered examinations is not on its own a complete method for identifying issues with difficulty targeting or standard setting. Comparing thresholds at grade D might result in a different interpretation. Moreover, issues with difficulty targeting or standard setting might not be reflected in reduced or reversed differences between the Foundation and Higher thresholds. The method of comparing C grade thresholds is recommended because it is straightforward, easy to automate and can then be done routinely as part of a wider monitoring system.

When a reduced or reversed difference between Foundation tier and Higher tier C-thresholds is detected, it is important to understand what has caused it. If items did not function as intended and an examination was harder or easier than it should have been, it is appropriate to set lower or higher thresholds respectively to compensate. Thus an unexpected difference between Foundation and Higher tier C-thresholds does not imply that either threshold was wrong or that the standards applied were not comparable; it can simply reflect the fact that the difficulty of one of the examinations was not optimal for its tier. Once this has been detected by means of the simple techniques described in this article, further investigations can take place to identify improvements for future examinations.

Acknowledgements
I would like to thank my colleagues Nick Raikes, Tom Bramley, Beth Black and Mike Forster for their advice.

References

An investigation on the impact of GCSE modularisation on A level uptake and performance

Carmen L. Vidal Rodeiro Research Division

Background of the study

Over the past few years modular assessment has been gaining popularity in England, particularly in large scale assessments such as the General Certificates of Secondary Education (GCSEs), which are taken by the majority of 14–16 year olds. Instead of being assessed at the end of a two-year course by following a linear syllabus, GCSE modular courses allow the assessment to take place in specified sessions in both the first and second years of the course. When multiple assessment paths exist for the same subject, it is left to individual schools to decide whether the assessment should be modular or whether candidates should enter for a linear examination.

However, it has recently been suggested that these modular assessments led to changes in learning opportunities and in the interaction between learning and assessment. In particular, modular assessment has been criticised for leading to fragmentation of learning and to a lack of coherence in the learning experience, endangering what is called synoptic understanding (Hayward and McNicholl, 2007), as students have little time for reflection, skill development and consolidation of learning. Furthermore, modular assessment might not provide opportunities for deep learning and it might, instead, encourage a climate of cramming (Priestley, 2003). In addition, the increased assessment load can lead children to spend more time revising for the next exam, rather than simply benefiting from learning (Hodgson and
Spours, 2001). Finally, there is the view that the possibility of re-sitting modules may be lowering examination standards (De Wall, 2009), and that 'teaching to the test' time is heightened at the expense of deeper learning or enrichment activities (Thomson, 1988; Poon Scott, 2010).

In the context of the English Post-14 education framework there has also been a lack of public and teacher confidence in other aspects of modularisation. In particular, teachers at schools and at colleges are concerned that modular courses are not a good foundation for advanced study (e.g. A levels, qualifications taken by students at age 18) and there is the concern that students will learn a particular part of the course and then forget it. In fact, in a research study carried out to investigate the effects of modular assessment at GCSE level, Vidal Rodeiro and Nádas (2010) found that GCSE teachers doubted whether modular GCSEs provide a good preparation for A level studies:

*The only thing that would worry me with the modular system is ... [students] don’t look at it again and so I know a lot of students do that in December of Year 10. If they decide to go on to do A level and they haven't looked at material for a year and a half, I think that is a definite disadvantage.*

On the same lines, researchers, teachers and policy makers showed concerns about modularisation at GCSE leading to a significant dropout in the first year of advance studies (Institute of Education, 2010).

Another example of the concerns about modularisation, although in a higher education context, appears in Tan (1992). He found a profound negative impact on medics' level of understanding of physiology taught and assessed in a modular structure because students adopted a surface approach in order to pass their exams, rather than a high-level conceptual development of understanding the subject. As a consequence, they had difficulty in relating theory to practice later.

Some of the above issues might have had the current Secretary of State for Education to say, in November 2010, that:

*We want to get rid of modularisation of GCSE. Instead of GCSEs being split into bite-sized elements we think it is important that at the end of the GCSE course the student should be examined on everything they have learnt at one time. We will have fewer exams but a concentration on a more rigorous approach at age 16 (BBC News, 2010).*

Furthermore, in June 2011, the Secretary of State for Education stated again that the present system prevented students gaining a deep and rounded knowledge of subjects and was forcing England down international league tables, and he added that GCSEs were losing rigour by being chopped into bite size chunks (Davis, 2011).

This study set out to investigate whether different assessment routes (linear vs. modular) equipped students equally for further study. The focus was on the impact of the GCSE assessment route on the uptake and performance in three A level subjects: English, mathematics and ICT.

### Data and methods

#### Data

A range of GCSE and A level subjects were selected for this research. The subjects were chosen because they cover a range of curriculum areas and because their entries were reasonably high.

At GCSE level, three subjects offered by the OCR awarding body were selected: English, mathematics and ICT. At the time this research was carried out, OCR GCSEs in English and ICT were organised into modules (or units) which could either all be taken at the end of the course in a linear fashion or could be taken in different sessions throughout the two-year course (see OCR, 2003, and OCR, 2004, respectively, for a detailed description of the assessment in these subjects). However, in mathematics, OCR offered two separate specifications which were identical in content but different in structure. GCSE mathematics C (OCR, 2007) was a unitised specification and GCSE mathematics A (OCR, 2006) was a linear specification. At A level, subjects in the same knowledge area as the GCSE subjects were selected (see Table 1).

<table>
<thead>
<tr>
<th>Table 1: GCSE and A level subjects included in this study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GCSE (OCR only)</strong></td>
</tr>
<tr>
<td>Mathematics</td>
</tr>
<tr>
<td>English</td>
</tr>
<tr>
<td>ICT / Computing studies</td>
</tr>
</tbody>
</table>

Only candidates aged 15 at the start of the two-year GCSE course in September 2006 and who certificated in June 2008 taking the necessary units to do so in 2007 or 2008 sessions, were included in this research. This restriction was made in an attempt to mirror a typical GCSE cohort.

GCSE results in the three subjects, both at specification level and at unit level and for all examination sessions, were obtained from OCR’s examinations processing system. GCSEs are graded on an eight-point scale (A*, A, B, C, D, E, F and G) and those who fail to reach the minimum standard for grade G are recorded as U (ungraded).

A level uptake and performance figures for the 2010 examination session for all English boards were obtained from the National Pupil Database, which is compiled by the Department for Education and contains individual-level information and attainment records for all students in schools within England. A levels are graded on a six-point scale (A*, A, B, C, D and E). Those who fail to reach the minimum standard for grade E are recorded as U (ungraded).

#### Methods

The candidates certificating in June 2008 in any of the OCR GCSE subjects listed in Table 1 were matched to the 2010 National Pupil Database, which contains all qualifications taken by candidates who were at the end of Key Stage 5 in 2010. In doing so, candidates were linked to the results of the A level subjects in which they certificated in 2010 (see Table 2).

<table>
<thead>
<tr>
<th>Table 2: OCR GCSE candidates and their progression to A level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subject</strong></td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>Mathematics</td>
</tr>
<tr>
<td>English</td>
</tr>
<tr>
<td>ICT</td>
</tr>
</tbody>
</table>

Descriptive statistics were used to investigate uptake and performance patterns in each A level subject for both GCSE assessment routes. However, an assessment route at GCSE might exhibit higher progression...
to A level or better performance at A level simply because it had a more
able candidature. To resolve this limitation, the descriptive analyses were
followed by a multilevel logistic regression analysis which accounted for
the ability of the students.

Logistic regression is a type of regression analysis that is used
when the dependent variable or outcome is a dichotomous variable
(i.e. it takes only two values, which usually represent the occurrence or
non-occurrence of some event) and the independent variables are
continuous, categorical, or both. It is used to predict the probability that
the event of interest will occur as a function of the independent variables
(see, for example, Hosmer and Lemeshow, 2000). The multilevel model
was proposed due to the hierarchical or clustered structure of the data
(as students were grouped within schools). If we failed to recognise this
hierarchical structure, the standard errors of the regression coefficients
would be underestimated, leading to an overstatement of the statistical
significance.

In the following, the multilevel logistic regression model fitted to
answer the question: ‘What is the impact of the GCSE assessment route on
A level uptake?’ is described in detail. With small variations the model can
be applied to answer the question: ‘What is the impact of the GCSE
assessment route on A level performance?’.

The dependent variable was the uptake of a subject at A level with the
variable taking the value 1 if the candidate entered for an examination in
the subject and 0 otherwise. The independent or explanatory variables were:
gender, GCSE assessment route, attainment at GCSE and school type. These variables were categorical with the exception of the
attainment at GCSE which was treated as a continuous variable.

Interaction terms between gender and assessment route and between
attainment and assessment route were also included. Schools were
classified into the following groups: comprehensive, academy, grammar,
independent and secondary modern.

The categorical independent variables had a baseline category with
which all other categories in the variable were compared. ‘Male’ was
taken as the reference for gender, ‘comprehensive’ for school type and
‘modular’ for GCSE assessment route.

The formal representation of the model was:

$$\log \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 \text{Gender}_{ij} + \beta_2 \text{Assessment Route}_{ij} + \beta_3 (\text{Gender} \times \text{Assessment Route})_{ij} + \beta_4 \text{Attainment}_{ij} + \beta_5 (\text{Attainment} \times \text{Assessment Route})_{ij} + \beta_6 \text{School Type}_{ij} + u_j$$

where $p_i$ was the probability of student $i$ in school $j$ taking the subject
at A level, $\beta_0$ to $\beta_4$ were the regression coefficients or fixed effects and $u_j$
was a random variable at school level which followed a normal
distribution with mean zero and therefore it was sufficient to estimate its
variance.

In the first instance, main effects and interaction terms were included
in all models. However, for simplicity, the interaction terms were not
included in the final models if they were found not to be statistically
significant.

The results of the multilevel logistic regressions, that is, the regression
coefficients, are reported in the full report of this research (Vidal Rodeiro,
2011). In this article, estimates of the probabilities of taking the subject
at A level by the candidates’ gender and ability in the subject, the GCSE
assessment route and the school attended at GCSE are reported instead.
This is an alternative representation of the data which offers an easier
interpretation.

In order to investigate the impact of the GCSE assessment route on
A level performance, two levels of A level attainment were considered:
achieving grade E or above and achieving grade A or above. In this
instance, the dependent variable for the multilevel logistic regression
models was the presence of a grade (e.g. grade A or above) in the A level
subject under consideration with the variable taking the value 1 if the
candidate obtained the grade and 0 otherwise. The independent variables
were, as before, gender, GCSE assessment route, attainment at GCSE and
school type.

**Results**

**Uptake of A level subjects**

**Mathematics**

In June 2008, 90732 candidates were awarded a GCSE in mathematics
by the OCR awarding body. Among those, 6.81% carried on to study an
A level in mathematics, certificating in June 2010. Table 3 shows that the
modular GCSE mathematics specification was more popular than the
linear one (the percentage of candidates following the modular route was
64.95%; this compares with 35.05% of the candidates following a linear
route). Very similar percentages of candidates in each GCSE assessment
route entered for an A level in mathematics.

<table>
<thead>
<tr>
<th>GCSE assessment route</th>
<th>GCSE %</th>
<th>A level in mathematics %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>31799</td>
<td>35.05</td>
</tr>
<tr>
<td></td>
<td>2200</td>
<td>6.92</td>
</tr>
<tr>
<td>Modular</td>
<td>58933</td>
<td>64.95</td>
</tr>
<tr>
<td></td>
<td>3976</td>
<td>6.75</td>
</tr>
<tr>
<td>All</td>
<td>90732</td>
<td>6176</td>
</tr>
<tr>
<td></td>
<td>6.81</td>
<td></td>
</tr>
</tbody>
</table>

The proportions of GCSE mathematics candidates who studied for an
A level in mathematics, by GCSE grade, are displayed in Table 4. This table
highlights that, among candidates who obtained grade A* in their
mathematics GCSE, higher percentages of modular candidates than linear
candidates continued to study mathematics at A level.

The probability of taking an A level in mathematics was modelled as a
function of the grade and assessment route in GCSE mathematics, the
candidates’ gender and the type of school where the GCSE was obtained.
The results of these analyses show that, once the attainment in
mathematics at GCSE was taken into account, candidates following a
linear assessment route at GCSE were more likely to continue to study
mathematics at A level, independently of the type of school they were in.

Figure 1 shows the probability of taking mathematics at A level for
students in comprehensive schools by the GCSE assessment route and
the grade in GCSE mathematics. In particular, Figure 1(a) shows that a
boy with a grade A* in GCSE mathematics (8 points) in a comprehensive
school had a probability of 0.88 of taking mathematics at A level if he
took GCSE mathematics in a linear fashion, and a probability of 0.86,
if he took the GCSE in a modular fashion. These differences, although
statistically significant, were small.

The differences between the GCSE assessment routes in other types of
schools were fairly similar to those in comprehensive schools and are
English

In June 2008, 45,911 candidates obtained a GCSE in English awarded by OCR. Among those, 6.15% carried on to study English literature at A level and certificated in June 2010. Table 5 shows that GCSE English was taken mainly in a linear fashion rather than in a modular fashion (83.00% vs. 17.00%) and that higher percentages of linear candidates than modular candidates pursued an A level in English literature.

The proportions of GCSE English candidates who studied for an A level in English literature, by GCSE grade, is displayed in Table 6. Percentages of modular and linear candidates at each grade entering for an A level in English literature were fairly similar.

1. GCSE grades below C were not considered in this analysis as only a very small proportion of candidates with such grades go on to study A levels (Bell and Emery, 2007).

2. Points were assigned to grades in each subject (A*=8, A=7, B=6, C=5).

<table>
<thead>
<tr>
<th>GCSE Grade</th>
<th>GCSE candidates</th>
<th>Candidates going onto A level</th>
<th>% going onto A level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Modular</td>
<td>Linear</td>
<td>All</td>
</tr>
<tr>
<td>A*</td>
<td>2747</td>
<td>1588</td>
<td>4335</td>
</tr>
<tr>
<td>A</td>
<td>6416</td>
<td>3692</td>
<td>10108</td>
</tr>
<tr>
<td>B</td>
<td>9274</td>
<td>4904</td>
<td>14078</td>
</tr>
<tr>
<td>C</td>
<td>14749</td>
<td>8446</td>
<td>23195</td>
</tr>
</tbody>
</table>

Figure 1: Probability of taking A level in mathematics by GCSE assessment route

<table>
<thead>
<tr>
<th>GCSE assessment route</th>
<th>GCSE</th>
<th>A level in English literature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Candidates</td>
<td>%</td>
</tr>
<tr>
<td>Linear</td>
<td>38107</td>
<td>83.00</td>
</tr>
<tr>
<td>Modular</td>
<td>7804</td>
<td>17.00</td>
</tr>
<tr>
<td>All</td>
<td>45911</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: A level uptake of linear/modular English GCSE students by GCSE grade

<table>
<thead>
<tr>
<th>GCSE Grade</th>
<th>GCSE candidates</th>
<th>Candidates going onto A level</th>
<th>% going onto A level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Modular</td>
<td>Linear</td>
<td>All</td>
</tr>
<tr>
<td>A*</td>
<td>269</td>
<td>2981</td>
<td>3250</td>
</tr>
<tr>
<td>A</td>
<td>1014</td>
<td>7289</td>
<td>8303</td>
</tr>
<tr>
<td>B</td>
<td>1809</td>
<td>9551</td>
<td>11360</td>
</tr>
<tr>
<td>C</td>
<td>1956</td>
<td>8253</td>
<td>10209</td>
</tr>
</tbody>
</table>
The probability of taking an A level in English literature was modelled as a function of the grade and assessment route in GCSE English, the candidates’ gender and the type of school where the GCSE was obtained. The results of these analyses show that, once the attainment in English at GCSE was taken into account, there were no statistically significant differences in the probability of taking an A level in English literature between the group of students who took English GCSE in a linear fashion and those who did so in a modular way.

ICT

In June 2008 there were 15262 candidates who obtained a GCSE in ICT awarded by OCR. Among those, 5.85% continued to study either ICT or computing studies at A level and certificated in June 2010.

Table 7: A level uptake of linear/modular ICT GCSE students

<table>
<thead>
<tr>
<th>GCSE assessment route</th>
<th>GCSE candidates</th>
<th>A level candidates</th>
<th>% going onto A level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>11789</td>
<td>633</td>
<td>5.37</td>
</tr>
<tr>
<td>Modular</td>
<td>3473</td>
<td>260</td>
<td>7.49</td>
</tr>
<tr>
<td>All</td>
<td>15262</td>
<td>893</td>
<td>5.85</td>
</tr>
</tbody>
</table>

Table 8: A level uptake of linear/modular ICT GCSE students by GCSE grade

<table>
<thead>
<tr>
<th>GCSE Grade</th>
<th>GCSE candidates</th>
<th>Candidates going onto A level</th>
<th>% going onto A level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Modular</td>
<td>Linear</td>
<td>All</td>
</tr>
<tr>
<td>A*</td>
<td>89</td>
<td>360</td>
<td>449</td>
</tr>
<tr>
<td>A</td>
<td>442</td>
<td>1305</td>
<td>1747</td>
</tr>
<tr>
<td>B</td>
<td>860</td>
<td>2625</td>
<td>3485</td>
</tr>
<tr>
<td>C</td>
<td>875</td>
<td>2787</td>
<td>3662</td>
</tr>
</tbody>
</table>

ICT

In June 2008 there were 15262 candidates who obtained a GCSE in ICT awarded by OCR. Among those, 5.85% continued to study either ICT or computing studies at A level and certificated in June 2010.

Table 7 shows that GCSE ICT was taken mainly in a linear fashion rather than in a modular fashion (77.24% vs. 22.76%), and that higher percentages of modular GCSE students than linear ones pursued ICT or computing studies at A level.

The proportions of GCSE ICT candidates who studied ICT or computing studies at A level is displayed, by GCSE grade, in Table 8. At each GCSE grade, higher percentages of modular candidates than linear candidates continued to study ICT or computing studies at A level.

The results of the logistic regression analysis carried out to investigate if the probability of taking an A level in ICT or computing studies differed by GCSE attainment route show that linear students were less likely to take the A level than modular ones. This supports the percentages shown in Table 8.

Figure 2 shows the probability of taking ICT or computing studies at A level for students in a comprehensive school by GCSE assessment route and grade in GCSE ICT. The differences between both GCSE assessment routes in other types of schools were fairly similar to those in comprehensive schools and are presented in Vidal Rodeiro (2011).

Figure 2(a) shows that a boy with a grade A in GCSE ICT (7 points) in a comprehensive school had a probability of 0.23 of taking ICT or computing studies at A level if he took GCSE ICT in a linear fashion, and a probability of 0.28, if he took the GCSE in a modular fashion. As for mathematics, these differences, although statistically significant, were small.

The probability of taking an A level in English literature was modelled as a function of the grade and assessment route in GCSE English, the candidates’ gender and the type of school where the GCSE was obtained. The results of these analyses show that, once the attainment in English at GCSE was taken into account, there were no statistically significant differences in the probability of taking an A level in English literature between the group of students who took English GCSE in a linear fashion and those who did so in a modular way.

ICT

In June 2008 there were 15262 candidates who obtained a GCSE in ICT awarded by OCR. Among those, 5.85% continued to study either ICT or computing studies at A level and certificated in June 2010.

Table 7 shows that GCSE ICT was taken mainly in a linear fashion rather than in a modular fashion (77.24% vs. 22.76%), and that higher percentages of modular GCSE students than linear ones pursued ICT or computing studies at A level.

The proportions of GCSE ICT candidates who studied ICT or computing studies at A level is displayed, by GCSE grade, in Table 8. At each GCSE grade, higher percentages of modular candidates than linear candidates continued to study ICT or computing studies at A level.

The results of the logistic regression analysis carried out to investigate if the probability of taking an A level in ICT or computing studies differed by GCSE attainment route show that linear students were less likely to take the A level than modular ones. This supports the percentages shown in Table 8.

Figure 2 shows the probability of taking ICT or computing studies at A level for students in a comprehensive school by GCSE assessment route and grade in GCSE ICT. The differences between both GCSE assessment routes in other types of schools were fairly similar to those in comprehensive schools and are presented in Vidal Rodeiro (2011).

Figure 2(a) shows that a boy with a grade A in GCSE ICT (7 points) in a comprehensive school had a probability of 0.23 of taking ICT or computing studies at A level if he took GCSE ICT in a linear fashion, and a probability of 0.28, if he took the GCSE in a modular fashion. As for mathematics, these differences, although statistically significant, were small.
Performance in A level subjects

Mathematics

Figure 3 presents the grade distribution for A level mathematics by GCSE mathematics assessment route. The two groups of students (linear or modular) differ significantly in the proportions obtaining each of the different A level grades. A larger percentage of linear GCSE candidates obtained grades A* and A at A level than modular GCSE candidates.

The assessment route at GCSE was found to have a statistically significant effect on the probability of obtaining at least grade A and at least grade E in mathematics at A level, after controlling for the grade obtained in GCSE mathematics. In particular, the probability of obtaining at least grades A or E in A level mathematics was higher for those candidates who followed a linear route in the subject at GCSE (Figures 4 and 5). Figure 4(a) shows, in particular, that a boy with a grade A* in GCSE mathematics (8 points) had a probability of 0.79 of obtaining grade A or above in mathematics A level if he took the GCSE in a linear fashion, and a probability of 0.66, if he took the GCSE in a modular fashion.

Figure 4: Probability of obtaining at least grade A in A level mathematics by GCSE assessment route

(a) Grade A or above – Boys
- linear assessment at GCSE
- modular assessment at GCSE

Figure 5: Probability of obtaining at least grade E in A level mathematics by GCSE assessment route

(a) Grade E or above – Boys
- linear assessment at GCSE
- modular assessment at GCSE

(a) Grade E or above – Girls
- linear assessment at GCSE
- modular assessment at GCSE
Although these analyses show that students following a linear route had a higher probability of obtaining certain grades at A level mathematics than those following a modular one, it should be noted that this does not imply that a causal relationship exists between GCSE assessment route and A level performance. There might be other factors, such as students’ motivation or students’ background that could be affecting performance and have not been taken into account (Vidal Rodeiro and Bell, 2007; Vidal Rodeiro, Emery and Bell, 2011).

**English**

In A level English literature, the two groups of students (linear or modular) differ significantly in the proportions obtaining each of the different A level grades and, in particular, higher percentages of linear candidates than modular candidates obtained grades A* and A in the A level subject (see Figure 6).

After modelling the probability of obtaining a certain grade at A level in English literature controlling for the attainment at GCSE, it was found that there were no significant differences in performance at any given grade in English literature between the group of students who took GCSE English in a linear fashion and those who did so in a modular way.

**ICT**

Figure 7, which presents the grade distribution for A level ICT/computing studies by GCSE assessment route, shows that the two groups of students (linear vs. modular) did not differ significantly in the proportions across the different grades.

After modelling the probability of obtaining a certain grade in A level ICT/computing studies, it was found that there were no significant differences in performance at any given grade between the group of students who took ICT GCSE in a linear fashion and those who did so in a modular way, once attainment at GCSE was taken into account.

Conclusions and discussion

This study set out to investigate whether different assessment routes (linear vs. modular) equipped students equally for further study and presented some evidence that uptake and performance in A level subjects differed depending on the assessment route of the subject at GCSE. However, in considering the conclusions of this research, some limitations with the current study must be noted.

First, there were some limitations regarding the data used for this research. The linking between candidates with OCR GCSE qualifications in 2008 and candidates recorded in the Key Stage 5 extract of the 2010 National Pupil Database was carried out using a unique pupil number common to both datasets. However, matching in this way was not perfect and it would have been impossible to achieve a 100% matching rate. Therefore, some candidates who progressed to A level (or to A level in a specific subject) might not have been included in the analyses. In addition, the selection of subjects for analysis was limited by the OCR GCSE subjects that were available, in 2008, in a unitised version (in order to follow up linear and modular candidates to A level in 2010). In 2008, OCR offered unitised GCSE specifications only in English, ICT, mathematics and the sciences.

Secondly, the research assumed comparability among A level specifications across the different awarding bodies in England in the subjects under consideration (that is, two specifications are comparable if candidates with given characteristics who demonstrate the same level on some other measure of attainment obtain the same grade).

Finally, the statistical analysis techniques (e.g. multilevel logistic regressions) applied in this research identify associative, not causal, relationships. In particular, although these analyses showed that uptake and performance in A level subjects differed depending on the route (modular vs. linear), it should be noted that this does not imply that a causal relationship exists between GCSE assessment route and the issues studied at A level. There might be other factors, such as students’ motivation or students’ background that could be affecting uptake or performance and have not been taken into account (Vidal Rodeiro and Bell, 2007; Vidal Rodeiro, Emery and Bell, 2011).

Despite the above limitations, the results presented in this report have provided evidence to show that uptake and performance at A level differed depending on the assessment route (modular vs. linear) at GCSE.
Table 9 in this section shows a summary of the results from the multilevel logistic regression analyses carried out in this study. In particular, Table 9 shows that:

- Students following a linear assessment route in GCSE mathematics were more likely to continue to study mathematics at A level than those who followed a modular route. Conversely, linear students were less likely to progress to A level in ICT. There were no differences between linear and modular GCSE students in the uptake of English literature at A level.

- Once attainment at GCSE was taken into account, the assessment route at GCSE only affected performance in A level mathematics, with students following a linear route in mathematics at GCSE performing better at A level than those following a modular one. Even when statistically significant, it should be noted that the differences at A level between GCSE modular and linear candidates in the two areas looked at in the research were very small. Furthermore, no GCSE assessment route offered consistently the best outcomes (i.e. higher uptake and better performance). Therefore, the claims about modular GCSEs not providing a good preparation for A level study do not seem to have a very strong base and should be investigated further. It should also be noted, as mentioned above, that there might be other factors that do not relate to the nature of the syllabuses (e.g. motivation, teachers' experience of the assessment route, etc.) that might be different in the different GCSE routes and could have an impact on further study.

Table 9: Summary of results

<table>
<thead>
<tr>
<th>A level subject</th>
<th>GCSE subject</th>
<th>Uptake at A level</th>
<th>Performance at A level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics</td>
<td>Mathematics</td>
<td>GCSE linear</td>
<td>GCSE linear candidates were more likely to achieve at least grade A and at least grade E at A level</td>
</tr>
<tr>
<td>English</td>
<td>English</td>
<td>No difference</td>
<td>No difference</td>
</tr>
<tr>
<td>ICT/computing</td>
<td>ICT</td>
<td>GCSE modular</td>
<td>No difference</td>
</tr>
<tr>
<td>studies</td>
<td></td>
<td>candidates were more likely to take the A level</td>
<td></td>
</tr>
</tbody>
</table>

As stated above, only in one of the A level subjects investigated in this study (mathematics), were linear candidates more likely to perform better than their modular counterparts. It should be noted that among the GCSE subjects considered in this report, mathematics was the only one where there was a linear specification on offer (the other GCSE subjects were unitised and students could take all necessary units for certification in one session or in different sessions throughout the course). Therefore, this research has particularly shown that only when the GCSE specification was linear in design, was there an impact on A level performance.

The outcomes of this research provide evidence to inform key issues in an area of assessment which is currently under the spotlight as debate continues about the balance of advantages and disadvantages of these two different types of assessment.

References


Piloting a method for comparing the demand of vocational qualifications with general qualifications

Jackie Greatorex and Hannah Shiell Research Division

Introduction

The demand of vocational and general qualifications receives much attention. One view is that centres offer, and learners take, vocational qualifications as a purportedly less demanding route to good grades and to boost centres’ performance in league tables (Davis, 2011; Stewart 2010; Paton, 2008, 2010a, b, c). The Wolf Review and a government consultation considered these issues (Wolf, 2011, DfE, 2011a and b). The outcome was that some vocational qualifications remain in league tables (DfE, 2011c). Comparability research helps evaluate the aforementioned view by providing robust information about demand.

To undertake such research an instrument is needed. This article reports the pilot of such an instrument. A glossary is provided.

Greatorex and Rushton (2010) and Novaković and Greatorex (2011) reviewed several comparability studies to determine how best to conduct a comparison of demands. All the reviewed studies used a research instrument to gather expert decisions about demands. The reviews indicated that a research instrument should:

- Gather expert decisions in the form of paired comparisons by instructing experts to decide which unit is more demanding. This is repeated for many pairs.
- Cover a variety of domains (areas of knowledge) such as the affective, cognitive and psychomotor domains.

Therefore these two characteristics were incorporated in the instrument piloted here.

The researchers were tasked with investigating methods of comparing general qualifications with vocational qualifications at level 2. Reading a variety of OCR level 2 specifications illustrated that they included knowledge, skills and understanding from five domains (the affective, cognitive, interpersonal, metacognitive and psychomotor domains).

Further details about the domains are in Figure 1. These domains were included in the instrument so it should be suitable for use with different types of qualifications.

Domains do not indicate what is more and less demanding; this information is gained from taxonomies. A taxonomy is: “a classification system that establishes the hierarchy of the parts to the whole” (Hauenstein, 1998, p.2).

A taxonomy for each domain was chosen from existing literature:

- Affective (Hauenstein, 1998)
- Cognitive (Hauenstein, 1998)
- Interpersonal (Rackham and Morgan, 1977)
- Metacognitive (Howell and Caros, 2006)
- Psychomotor (Hauenstein, 1998).

The demands instrument is included in Appendix A.

Comparability studies draw heavily on expert judgement and research about expert judgement has proved fruitful for other areas of assessment practice. For instance, Laming (2004) found that all judgements are relative, that is, they are comparisons of one thing with another. This was used to argue for comparability study methods that ask experts to make relative judgements (paired comparisons/rank ordering). For example, in Kimbell et al. (2007) experts were presented with many pairs of scripts and for each pair they decided which script was better. Pairs were constructed from combinations of current and previous exam scripts. The decisions were analysed to determine grade boundaries comparable
with previous grade boundaries. Given that expert judgement research has a track record of being useful, it was decided to further investigate comparability judgement in the present study.

Research questions

Two research questions were investigated:

1. Is the demands instrument appropriate for use in research studies? (i.e. did the results make comparisons between the demand of different types of units?)
2. How did experts judge which units were more demanding?

Judgement questionnaire

A judgement questionnaire was designed to evaluate whether the demands instrument was usable and to investigate how experts judged which units were more demanding. The rationale for each section of the questionnaire is detailed below. The questions from the questionnaire together with the response options are given later in the article.

Instruments are appropriate for use in research if they produce credible results. Previous comparability studies recruited senior assessors, on the basis that their expertise lent credibility to the results and their experience underpinned judgement of demand. (For further details see Elliott and Creatorex (2002) and Adams (2007).) Laming (2004) explains that judgements are heavily influenced by experience. Therefore it is important to know what experience participants thought they used to judge demand. How experience related to judging demand was addressed by Question 1 in the questionnaire.

The experts were instructed to use a concept of typical level 2 learners to judge demand. They cannot follow this instruction if they have no such concept. Therefore Question 2 asked the experts to share their concept of a typical level 2 learner. Additionally, in the interests of transparency it is important to know the basis for judgements. Research results are invalid if experts judge using invalid strategies. Therefore Question 3 investigated judgement strategies, and whether invalid strategies were invoked.

It was expected that experts would find it manageable to hold a concept of typical learners in mind. Novaković (2008) found this to be the case in other assessment situations. Question 4 addressed these issues.

The three anticipated problems of using the demands instrument were that experts were:

1. Judging using concepts other than typical level 2 learners. Experts conceptualised incorrect groups of learners in other assessment situations for example by putting themselves in the place of particular learners (Novaković, 2008) or thinking about familiar learners (Skorupski and Hambleton, 2005).
3. Experiencing concept drift, i.e. using different concepts of a typical level 2 learner at different points in the study. Ricker (2006) explained that a limitation of some assessment situations is experts experiencing concept drift.

Questions 5, 6 and 7 addressed these points.

Method

Units

Four cognate Health and Social Care level 2 units were selected as listed in Table 1.

Table 1: Type of qualifications and Health and Social Care qualification from which the units were sourced

<table>
<thead>
<tr>
<th>Type of qualification</th>
<th>Health and Social Care Qualifications</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>VQ</td>
<td>NVQ</td>
<td>NVQ1</td>
</tr>
<tr>
<td>GQ</td>
<td>GCSE</td>
<td>GCSE1</td>
</tr>
<tr>
<td>VQ</td>
<td>NVQ</td>
<td>NVQ2</td>
</tr>
<tr>
<td>GQ</td>
<td>GCSE</td>
<td>GCSE2</td>
</tr>
</tbody>
</table>

The two NVQ units selected were from the same NVQ.

Experts

Four research participants were recruited. They were each a team leader or assistant external verifier or higher for at least one of the qualifications in the study¹ and were all recommended by OCR.

Materials

The experts needed to be familiar with specifications of the units in order to participate in the research. As specifications are substantial documents, which are time-consuming to read, extracts were used rather than whole specifications. The extracts included:

- Aims of the specification
- Assessment objectives of the unit
- Unit content
- Assessment structure
- Information about guided learning hours or assessment time
- Grade/performance descriptors
- Teaching arrangements

The experts were also provided with a document containing the following materials:

- An introduction, instructions, and descriptions of domains and taxonomies
- The demands instrument
- The judgement questionnaire

The demands instrument required experts to compare pairs of units and decide which unit was more demanding for each domain. The experts were also asked to explain their decisions. All possible pairs were compared in this way.

1. This information was taken from OCR records.
Procedure
The experts individually:
- Read the definitions of demands, domain and taxonomy
- Read the specification extracts noting instances of affective, cognitive, interpersonal, metacognitive and psychomotor demands
- Completed the demands instrument and the judgement questionnaire

The experts were sent the materials, which were completed remotely and returned to the Research Division in hard copy.

Analysis
Demands instrument data
The data were analysed in two ways:
1. The level of consensus between experts about whether a unit was more demanding within a domain was calculated. It was noted when all four experts were in consensus that a particular unit in a pair was more demanding in a given domain.
2. The frequency a unit was judged more demanding in a domain was used to rank all four units from the most to the least demanding.

Point 1 focuses on comparing a pair of units, whereas point 2 focuses on comparing all four units.

Judgement questionnaire data
The frequency of responses to closed questions was calculated. The responses to open questions were divided into sections of text and each section of text was categorised. The frequency of experts whose response was classified in each category was calculated. The experts’ explanations for decisions are not reported here because this is outside the scope of the present study. If there is a consensus amongst experts that there is a difference in demand, the awarding body might decide to change a specification. In such cases the explanations might provide details to guide the changes.

Results
Demands instrument
Consensus amongst experts
Experts individually decided which unit was more demanding for each pair of units. The results are presented in Table 2 which shows the frequency of experts who judged one unit to be the more demanding in a pair for a particular domain. A unit should only be considered to be more demanding when there was a consensus amongst all four experts. The consensus agreements are shaded in the table. For example, four experts judged NVQ1 to be more demanding than GCSE1 in the affective domain, therefore a consensus was reached. Where a cell contains 1, 2 or 3, it indicates there was no consensus on which unit was more demanding.

Table 2 indicates consensus amongst all four experts that:
- An NVQ unit was more demanding than a GCSE unit in six pairs
- A GCSE unit was more demanding than an NVQ unit in three pairs

Comparisons were made between units of the same type, as well as units of different types.

Table 2 Level of consensus between experts

<table>
<thead>
<tr>
<th>Paired comparison</th>
<th>Affective</th>
<th>Cognitive</th>
<th>Interpersonal</th>
<th>Metacognitive</th>
<th>Psychomotor</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVQ1 more demanding than GCSE1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>GCSE1 more demanding than NVQ1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>NVQ1 more demanding than NVQ2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>NVQ2 more demanding than NVQ1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>NVQ1 more demanding than GCSE2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>GCSE2 more demanding than NVQ1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>GCSE1 more demanding than NVQ2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>NVQ2 more demanding than GCSE1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>GCSE1 more demanding than GCSE2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>GCSE2 more demanding than GCSE1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>NVQ2 more demanding than GCSE2</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>GCSE2 more demanding than NVQ2</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Indicates all four experts were in consensus that the unit was the more demanding in the pair for a domain

Ranking four units from the most to the least demanding
Table 3 and Figure 2 provide a more holistic picture than that provided in Table 2.

Table 3 and Figure 2 show that for each domain:
- The frequency each unit was judged to be more demanding
- The ranking of units from the most to the least demanding

For instance, in the cognitive domain GCSE2 was judged the most demanding on nine occasions; GCSE1 was judged the most demanding on eight occasions; NVQ2 was judged the most demanding on six occasions; and NVQ1 was judged the most demanding just once. Therefore GCSE2 was ranked the most demanding in the cognitive domain followed by GCSE1, then NVQ2 and NVQ1 was the least demanding.

Whilst it is possible to rank the units in terms of demand the ranks must be treated with caution. Sometimes a unit is ranked higher (or lower) than another unit but the ‘differences’ in demand are best interpreted as a lack of consensus between experts about whether one unit is the more demanding. Nevertheless, they provide a way of viewing the results for all units at once.

Judgement questionnaire
Table 4 to Table 10 give the frequency of responses to the questions.
Table 4 shows that experts reported that they drew from a variety of experience to make their judgements, particularly GCSE2 experience. For sub-questions h to r there were some missing responses.
Experts generally thought it was easy or very easy to conceptualise particular groups of learners as indicated in Table 5. Table 6 shows all experts strongly agreed or agreed that they used the concept of typical level 2 learners to decide which unit was more demanding. Fewer experts strongly agreed or agreed they used other concepts to judge demand.

Table 7 shows that experts generally thought it was very easy or easy to use concepts of a group of learners to make judgements. Most importantly, three of the four experts reported it was very easy to use the concept of a typical level 2 learner to make decisions about which unit was the most demanding.

All experts strongly agreed or agreed that they always used the same concept of typical learners as indicated in Table 8. It also shows all experts strongly agreed or agreed they put themselves in the place of familiar GCSE2 learners and thought about what typical GCSE2 learners find more and less demanding. Experts’ responses about the other units were more varied. This reflects that experts reported drawing from more GCSE2 than GCSE1 or NVQ experience, as indicated in Table 4.

Table 9 shows all four experts strongly agreed or agreed that specifications from different types of qualifications can be meaningfully compared and writers incorporate the demands they intend learners to experience in specifications. Two experts’ responses suggested they thought demand can be judged from specifications, and two experts neither agreed nor disagreed on this issue. Three experts strongly agreed or agreed that:

- For most learners some content is more demanding than other content
- Experts can judge differences in demand between units from different types of qualifications

### Table 3: Frequency that units were judged more demanding, and ranking by demand

<table>
<thead>
<tr>
<th>Domain</th>
<th>Unit</th>
<th>Frequency unit judged more demanding</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affective</td>
<td>NVQ2</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>NVQ1</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>GCSE2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>GCSE1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Cognitive</td>
<td>GCSE2</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>GCSE1</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>NVQ2</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>NVQ1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Interpersonal</td>
<td>NVQ1=</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>GCSE2=</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>NVQ2</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>GCSE1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Metacognitive</td>
<td>GCSE1</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>NVQ1</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>NVQ2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>GCSE2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Psychomotor</td>
<td>NVQ1</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>GCSE1=</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>GCSE2=</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>NVQ2</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

Experts generally thought it was easy or very easy to conceptualise particular groups of learners as indicated in Table 5.

Table 6 shows all experts strongly agreed or agreed that they used the concept of typical level 2 learners to decide which unit was more demanding. Fewer experts strongly agreed or agreed they used other concepts of learners to judge demand.

Table 7 shows that experts generally thought it was very easy or easy to use concepts of a group of learners to make judgements. Most importantly, three of the four experts reported it was very easy to use the concept of a typical level 2 learner to make decisions about which unit was the most demanding.

All experts strongly agreed or agreed that they always used the same concept of typical learners as indicated in Table 8. It also shows all experts strongly agreed or agreed they put themselves in the place of familiar GCSE2 learners and thought about what typical GCSE2 learners find more and less demanding. Experts’ responses about the other units were more varied. This reflects that experts reported drawing from more GCSE2 than GCSE1 or NVQ experience, as indicated in Table 4.

Table 9 shows all four experts strongly agreed or agreed that specifications from different types of qualifications can be meaningfully compared and writers incorporate the demands they intend learners to experience in specifications. Two experts’ responses suggested they thought demand can be judged from specifications, and two experts neither agreed nor disagreed on this issue. Three experts strongly agreed or agreed that:

- For most learners some content is more demanding than other content
- Experts can judge differences in demand between units from different types of qualifications
Table 5: Frequency of responses to question 4

4) How easy/hard was it for you to conceptualise the following learners?

<table>
<thead>
<tr>
<th>Learner Type</th>
<th>Very Easy</th>
<th>Easy</th>
<th>Neither easy nor hard</th>
<th>Hard</th>
<th>Very Hard</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Typical level 2 learners</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(as described in question 3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. Familiar typical level 2 learners</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c. The majority of level 2 learners</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d. Average level 2 learners</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>e. Very able level 2 learners</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f. Less able level 2 learners</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6: Frequency of responses to question 5

5) To what extent do you agree with the following statements?

To decide which unit was more demanding I used my concept of...

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Neither agree nor disagree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Typical level 2 learners (as described in question 3)</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b. Familiar typical level 2 learners</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c. The majority of level 2 learners</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d. Average level 2 learners</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>e. Very able level 2 learners</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f. Less able level 2 learners</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 7: Frequency of responses to question 6

6) How hard/easy was it for you to use the following concepts to judge which was the most demanding unit using?

<table>
<thead>
<tr>
<th>Concept</th>
<th>Very Easy</th>
<th>Easy</th>
<th>Neither easy nor hard</th>
<th>Hard</th>
<th>Very Hard</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. A typical level 2 learner (as described in question 3)</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b. Familiar typical level 2 learners</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c. The majority of level 2 learners</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d. Average level 2 learners</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>e. Very able level 2 learners</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f. Less able level 2 learners</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 8: Frequency of responses to question 7

7) To what extent do you agree with the following statements?

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Neither agree nor disagree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. I always used the same concept of typical learners</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b. I put myself in the place of familiar GCSE2 learners</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c. I thought about what typical GCSE2 learners find more and less demanding</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d. I put myself in the place of familiar GCSE1 learners</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>e. I thought about what typical GCSE1 learners find more and less demanding</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f. I put myself in the place of familiar NVQ2 learners</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>g. I thought about what typical NVQ2 learners find more and less demanding</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>h. I put myself in the place of familiar NVQ1 learners</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>i. I thought about what typical NVQ1 learners find more and less demanding</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>j. I thought about what is more and less demanding for level 2 learners</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 9: Frequency of responses to question 8

8) To what extent do you agree with the following statements?

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Neither agree nor disagree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Specifications from different types of qualifications can be meaningfully compared in a comparability study</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b. When specification writers develop specifications they incorporate the demands they intend learners to experience</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c. Demands can only be judged from assessment tasks such as exam questions, not specifications</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d. Some content of specifications and associated activities is more demanding than others for the majority of learners</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>e. Experts can rank small numbers of specifications from the most to the least demanding</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f. Experts can judge differences in demand between units from different types of qualifications (e.g. general versus vocational qualifications)</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Qualitative responses

The experts explained their strategy for making comparisons. There were some strategies which were mentioned by more than two experts such as:

- Knowledge and understanding required – depth – progression throughout the unit to achieve the intended outcomes.

However, there were other strategies only mentioned by one expert, such as “Clarity of concepts/instructions.”

There are more than four strategies in Table 10 because some experts had more than one approach to making a decision.

Experts described several characteristics of typical level 2 learners, presented in Table 11. Some experts listed more than one characteristic. There were three characteristics each mentioned by two experts, for example, “have breadth but not depth to their knowledge, skills and application”. However, there were also an additional nine characteristics and each one was mentioned by only one expert, for example, “Learn by rote”.

In Table 10 and Table 11 direct quotes are presented in quotation marks but sometimes the expert views were summarised, in which case quotation marks are not used.

Table 10: Responses to question 2

2) Explain your strategy for deciding which unit was the most demanding

<table>
<thead>
<tr>
<th>Example/summary of comments</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>“I compared the units side by side looking at the content and the assessment requirements/guidance for each unit. Alongside looking at the requirements of the domain and how this linked to the skills for each. I then balanced each of the (unit content and domain) to decide which was the most demanding.”</td>
<td>2</td>
</tr>
<tr>
<td>“Use of particular verbs.”</td>
<td>2</td>
</tr>
<tr>
<td>“Knowledge and understanding required – depth – progression throughout the unit to achieve the intended outcomes.”</td>
<td>2</td>
</tr>
<tr>
<td>“Strategies to be used as recommended for assessment purposes.”</td>
<td>1</td>
</tr>
<tr>
<td>“Range of ways of applying knowledge content to test and recall.”</td>
<td>1</td>
</tr>
<tr>
<td>“Clarity of concepts/instructions.”</td>
<td>1</td>
</tr>
<tr>
<td>“Independent research/by candidates overall/holistic approach given to criteria.”</td>
<td>1</td>
</tr>
</tbody>
</table>

Discussion

The research questions of this study were:

1. Is the demands instrument appropriate for use in research studies?
2. How did experts judge which units were more demanding?

There are several limitations with comparability research in general (see Newton et al., 2007). The limitations of the present study include:

- The sample of experts was small (n=4).
- The sample of experts had more GCSE than NVQ experience, according to OCR records. Despite this, where the four experts were in consensus then an NVQ unit was more demanding twice as often as a GCSE unit. Therefore the experts were not biased in favour of their area of experience [GCSE] being more demanding. This is a pleasing result as Massey and Newbould (1977) and Coles and Matthews (1995) found senior examiners judged examinations from their area of experience to be more stringent than other qualifications and Pollitt and Elliott (2003) reported that some subsequent comparability studies were designed accordingly.

It is important to be mindful of these limitations, however, the study does offer useful insights into comparing demands.

1. Is the demands instrument appropriate for use in research studies?

Findings from the demands instrument

The demands instrument is appropriate for use in Cambridge Assessment comparability research of level 2 specifications because the results can be used to compare the demand of units. Evidence for this claim is that in 11 out of 30 pairs of units, there was a consensus between all four experts about which unit was more demanding. A consensus was not expected for all pairs, because units may sometimes be of similar demand.

The results can be used to compare units from different types of qualifications, as intended. For instance, all four experts were in consensus that:

- For six pairs an NVQ unit was more demanding than a GCSE unit
- For three pairs a GCSE unit was more demanding than an NVQ unit

Additionally, the results can be used to compare units from the same type of qualification. For instance, all four experts were in consensus that:

- For one pair, a GCSE unit was more demanding than the other GCSE unit
- For one pair, an NVQ unit was more demanding than the other NVQ unit
Findings from the judgement questionnaire

All the experts agreed or strongly agreed with some of the assumptions of the study. Therefore comparability studies about different types of qualifications have some credibility.

However, one expert did not agree or strongly agree that experts can do the required tasks. This is a cautionary note and suggests that if the demands instrument is used in further research some checks should be made on expert decisions, for example, checking whether one expert consistently disagreed with the others. If so, researchers could consider removing the expert as an outlier. Alternatively, they could check whether the expert panel includes a variety and balance of experience. If not, more decisions may be needed.

2. How did experts judge which units were more demanding?

Responses to the judgement questionnaire

Research instruments are suitable for research purposes if they produce valid results. The validity of results from the research instrument relies on expert judgement. If the experts followed the instructions then the results are valid. If the experts could not or did not follow the instruction then validity was compromised. These principles underpin the discussion below.

Experts’ experience

A variety of valid experiences are important in judging, including teaching, qualifications, and experience as an assessor and verifier.

Identifying the concept of a typical level 2 learner

There was some commonality and some individuality in experts’ concepts of typical level 2 learners. Diversity can be an advantage if it is more representative. On the other hand, experts sharing a concept can be interpreted as more reliable. The experts generally said it was very easy or easy to conceptualise various groups of level 2 learners, including typical level 2 learners. This was expected because experts find it easy to conceptualise groups of learners in other assessment situations (see Novaković, 2008). It is also encouraging that the experts generally found conceptualising typical level 2 learners manageable, because this is a key part of making judgements and following the instructions.

Using the concept of typical level 2 learners

As instructed, the experts used their concept of typical level 2 learners in their judgements and found it very easy or easy to do so. This finding is positive. However, experts also used and found it easy or very easy to use other concepts of groups of learners such as the "very able" and "less able". It is unclear whether this compromised their judgements. These results were unexpected because in other assessment situations, experts found it difficult to use concepts of groups of learners, for example, Bouriscot and Roberts (2006).

Concept drift

All the experts strongly agreed or agreed that they maintained the same concept of typical learners throughout the study. It is pleasing the experts thought they did not experience concept drift which is a limitation of some assessment procedures; for further details see Ricker (2006).

In summary, three problems were anticipated and the questionnaire results suggested only one of these problems occurred; that experts used concepts of non-typical learners. The validity of the research is limited because of this, but overall, the performance of the demands instrument was better than expected.

Conclusion

The demands instrument is considered suitable for use in Cambridge Assessment comparability research for comparing the demand of cognate units in five domains. Such research does not measure the size of differences in demand or the overall demand, and this is important to acknowledge whenever the demands instrument is used. Qualification standards are at specification level. Therefore, a difference in the demand between two units of two different qualifications may not be a cause for concern, so long as the overall demand of the specification is appropriate.

There is little research about the demand of specification content and the pilot suggests that it can be credibly conducted. The comparisons are useful in several areas:

- Qualification development – i.e. checking that draft units are of comparable demand to existing specifications.
- Comparing qualifications when it is not possible to compare learners’ performance. This can happen when performance is assessed by observing work-based practice and therefore there are no artefacts created by learners to evidence the quality of performance (Greatorex, 2011).

An area for further research is whether an on screen version of the demands instrument is appropriate for future comparability research at Cambridge Assessment. If so, this would be in line with many examiner activities tending to be undertaken on screen rather than on paper, for example, examination question writing.

Glossary

Cognate “Related or analogous in nature, character, or function.” (www.thefreedictionary.com/cognate)

Demanding The extent to which a specification is intended to be challenging for typical learners.

Demand(s) The level of knowledge, skills and understanding required of typical learners to successfully complete a specification. The requirements might be in the: Affective, Cognitive, Interpersonal, Metacognitive and Psychomotor domains. Demand is a relative term, it could be replaced with ‘relative demand’ throughout the article. But demand is used for the purposes of brevity.

Domain A domain is a “sphere of knowledge or intellectual activity”. (Hauenstein, 1998, 2)

Specification “The complete description – including optional and mandatory aspects – of the content, assessment arrangements and performance requirements for a qualification. A subject specification forms the basis of a course leading to an award or certificate.” (www.qcda.org.uk)

Taxonomy A taxonomy is defined as "a classification system that establishes the hierarchy of the parts to the whole." (Hauenstein, 1998, 2.) Each domain has its own taxonomy. Each taxonomy outlines what is more and less demanding in each domain. The taxonomies and domains are given in.
Unit The smallest part of a qualification for which learners can gain a certificate.

References


Appendix A: Extracts from the instructions to experts and demand research instrument

Instructions
Enclosed are four specification extracts:

1. NVQ1 2. GCSE1 3. NVQ2 4. GCSE2

Please read and familiarise yourself with the four extracts, and note information in each specification relating to the following five domains of knowledge: affective; cognitive; interpersonal; metacognitive; and psychomotor. The domains are defined below.

If you wish to make notes, please use a separate piece of paper, or write on the specification extracts.

[Figure 1 (shown on page 29) was provided here]

Demands research instrument

Introduction
The demands research instrument was developed to compare units in terms of the demand the specification intends to place on typical level 2 learners.

Instructions

- The specification demands research instrument presents you with further information related to the five domains introduced earlier (affective, cognitive, interpersonal, metacognitive and psychomotor). Each domain has its own taxonomy which describes the dispositions and abilities which are more and less demanding for learners within that domain.

- There are several levels within each taxonomy (with the exception of the interpersonal taxonomy). In each case level 1 represents the least demanding learner dispositions/abilities, and each higher level is progressively more demanding for learners. You are not required to allocate levels to the specification; the taxonomies and levels are included to aid understanding of the domains.

- Please read the descriptions of the domains and taxonomies and levels on the following pages. Then refer back to your specification extracts and decide which unit intends to place greatest demand on typical level 2 learners in each of the five domains. Use the response sheets to record your decisions.

Please note that:

- The most demanding unit may not be the same for all domains.

- Ties are not allowed; you must select one unit as the most demanding for each domain. If you struggled to make a distinction then when you have circled a unit please include a question mark to indicate that you found the judgement difficult.

- Each comparison must be made independently of the other judgements. In other words, do not deduce the outcome of one comparison from your previous decisions; it is acceptable for your responses to contain inconsistent decisions.

- You do not need to rely only on the information given explicitly in the specification extracts; some demands may be implicit in the specification, or you might know of them through experience.

- When making comparisons please remember that the taxonomies describe what is more and less demanding for learners.

- When making comparisons please remember to concentrate on the demand on typical learners; not more or less able learners, or learners who have special requirements under the Equality Act 2010. (Special considerations are dealt with under other awarding body work.)

You might want to refer to any notes you made on the specifications, about the affective, cognitive, interpersonal, metacognitive and psychomotor.

Descriptions of domains, taxonomies and levels

The next pages present more detailed explanations of the five domains of knowledge (see page [page number was provided here] for general domain descriptions and examples from specifications). The following extended descriptions include domains and their taxonomies for each domain, which describe dispositions and abilities which are more and less demanding for learners. Level 1 represents the least demanding level and the levels then become progressively more demanding. (The interpersonal domain is the exception and does not have levels.) The domains and taxonomies vary in style and structure because they have been developed from different sources.

Please read through the information about each domain and, using this information with the four specification extracts, decide which unit is more demanding. Please record your decision on the response sheets.

Please note that you are not expected to assign taxonomy levels to the specification extracts. The taxonomies are included to give you a better understanding of the domains.

[Adapted versions of the following taxonomies were provided here:

- Affective (Hauenstein, 1998)
- Cognitive (Hauenstein, 1998)
- Interpersonal (Rackham and Morgan, 1977)
- Metacognitive (Howell and Caros, 2006)
- Psychomotor (Hauenstein, 1998).

The affective, cognitive and psychomotor domain levels and descriptions were adapted from Hauenstein, A.D. (1998) *A Conceptual Framework for Educational Objectives, A Holistic Approach to Traditional Taxonomies*. University Press of America: Maryland, with the permission of the publisher.

The interpersonal taxonomy is adapted from Rackham, N. & Morgan, T. (1977) *Behaviour Analysis in Training*, McGraw-Hill: Maidenhead, with permission from Mr N Rackham.


This research instrument including the adapted taxonomies, domains and definitions are only to be used for the purposes of education research by Cambridge Assessment.]
The validity of teacher assessed Independent Research Reports contributing to Cambridge Pre-U Global Perspectives and Research

Jackie Greatorex Research Division and Stuart Shaw Cambridge International Examinations

Background

The Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999, p.9) frame test validity in terms of “the concept or characteristic that a test is designed to measure”. That is, the Standards reflect a construct-centred approach to test validity. This perspective draws on the view that the theoretical, underlying construct such as mathematical aptitude, represented by an observable test score is the foundation for evaluating a test. Thus “all test scores are viewed as measures of some construct” (AERA, APA, NCME, 1999, p.174). The claim of validity is that the test adequately reflects the constructs and can be used as basis for the inference of attainment or aptitude depending on the test purpose.

It is important, therefore, to establish that tests elicit performances that reflect intended constructs and that test developers and providers have recourse to a reasonably well-informed and coherent theoretical model underpinning the construct(s) of interest if they are to operationalise aspects of the construct(s) for practical assessment purposes. In reality, however, “Tests are imperfect measures of constructs because they either leave out something that should be included... or else include something that should be left out, or both.” [Messick, 1989, p.34]. If the construct(s) is not well defined and test tasks are inappropriate, then it will be difficult to support claims an awarding body wishes to make about usefulness of its assessments, including claims that tests do not suffer from construct under-representation and construct irrelevance (CI).

The focus of this research is construct irrelevance. Its working definition for this study is that CI occurs when irrelevant constructs

---

**Example Response Sheet**

For each row, circle the unit which is more demanding and indicate why the unit is more demanding using the appropriate domain and taxonomy information to explain your decision. If you struggled, include a question mark to indicate that you found the judgement difficult.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Unit</th>
<th>Why was the more demanding unit more demanding?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affective</td>
<td>NVQ1</td>
<td>GCSE1</td>
</tr>
<tr>
<td>Affective</td>
<td>NVQ1</td>
<td>NVQ2</td>
</tr>
<tr>
<td>Affective</td>
<td>NVQ1</td>
<td>GCSE2</td>
</tr>
<tr>
<td>Affective</td>
<td>GCSE1</td>
<td>NVQ2</td>
</tr>
<tr>
<td>Affective</td>
<td>GCSE1</td>
<td>GCSE2</td>
</tr>
<tr>
<td>Affective</td>
<td>NVQ2</td>
<td>GCSE2</td>
</tr>
</tbody>
</table>
systematically influence marks. A claim about construct relevance or irrelevance is an evidence-based judgement about the extent to which marks are interpreted as the skills and knowledge the assessment is intended to measure. The evidence includes the way in which the mark scheme is interpreted and applied.

Certain assessor behaviours are sources of construct irrelevance if they result in irrelevant constructs systematically influencing marks. It is important to stress that it is not the behaviours that necessarily result in CI – a behaviour is only a source of CI if marks are systematically influenced.

A review of literature relating to the assessment of school/college coursework or projects was undertaken to identify assessor behaviours that potentially affect assessment judgement. Research showed several behaviours, that is, assessors:

- compared a candidate’s performance with another candidate’s performance (Morgan, 1996; Crisp, 2010)
- expressed feelings towards a candidate (Vaughan, 1991; Crisp, 2010)
- laughed or noted amusement at a candidate or their performance (Vaughan, 1991; Crisp, 2010)
- predicted the quality of a candidate’s future performance (Barritt et al., 1986; Crisp 2010)
- expressed a view on their own assessment practice (Crisp, 2010)
- commented on a candidate’s characteristic such as skill/ability/gender (Barritt et al., 1986; Vaughan, 1991; Crisp 2010)
- used surface features of a candidate’s work in judgements (Morgan and Watson, 2002)
- estimated a candidate’s effort invested in the work (Crisp, 2010).

The Cambridge Pre-U Independent Research Report

The Cambridge Pre-U is an international post-16 qualification designed to prepare candidates to succeed at their university studies by fostering independent and self-directed styles of learning. It is administered by Cambridge International Examinations (CIE), and is available in a variety of subjects. The subjects include the Cambridge Pre-U Global Perspectives and Research (GPR). GPR comprises two components, Global Perspectives (GP) and the Independent Research Report (IRR).

They are designed to be taught as two successive one-year courses with IRR building on GP.

The focus of this article is the IRR. The syllabus, not the present article, is the authoritative reference document (Cambridge International Examinations, 2008).

The IRR is designed to help candidates develop the ability to learn critically and independently. The candidate chooses a topic, develops a title in the form of a research question and undertakes research to write an essay or report which is 4,500 to 5,000 words long. The report should have an introduction identifying and exploring terms and issues, as well as stating the scope of the research and why the research is worth undertaking. The report should also have a conclusion. The report should be readable to the candidates’ peers and the candidate should be able to explain it to a non-specialist in the subject area. The work should be independent and the candidate should engage intellectually with the sources of evidence (e.g. books, articles, the internet).

The candidate and teacher meet during the project to discuss the question and research. Teachers are active in determining the subject and scale of the report; they might conduct seminars/ workshops to discuss subject specific issues and approaches, later seminars might be used to share ideas. The teacher also assists candidates in identifying and locating sources of evidence, understanding and developing appropriate research methods and organisational skills. Once the candidate has chosen a research question there is minimum intervention by the teacher. There should be on-going opportunities in single or group tutorials to discuss progress. The role of the teacher is analogous to a higher education lecturer supervising an undergraduate dissertation.

IRR is assessed by the candidate’s teacher. The assessment includes the report, a short (five to ten minute) terminal interview (viva) to authenticate the work is the candidate’s, and the teacher’s observations, experience and records of the candidate’s progress in developing and producing the IRR.

Samples of marking are centre moderated by an internal moderator (IM) and externally moderated by CIE. External moderation checks the marking of the report using the mark scheme. (The IRR mark scheme is in the syllabus). It should be noted that part of the mark scheme is for assessing and internally moderating each candidate’s Knowledge and understanding of the research process and Communication as evidenced by teacher’s observations, experiences and records (AO1 and AO4a), see Table 1. That is, AO1 and AO4a are not externally moderated.

This situation arises as external moderators do not have the teacher’s and IM’s experience of the candidate to judge AO1 and AO4a. Teachers and IMs can make interim annotations and summary comments (hereafter annotations and comments) when marking the report. Outside of the IRR context markers make annotations and comments to explain decisions to others and support judgements during marking (Crisp and Johnson, 2007; Fowles, 2008). Therefore the IMs and EMs for IRR might find the annotations and comments useful. However, there is no requirement to make the annotations and comments in IRR marking.

### Table 1: Assessment Objectives AO1 and AO4a

<table>
<thead>
<tr>
<th>Assessment Objective</th>
<th>Task</th>
<th>Clarification</th>
</tr>
</thead>
<tbody>
<tr>
<td>AO1 Knowledge and understanding of the research process</td>
<td>Design, plan, manage and conduct own research project using techniques and methods appropriate to the subject discipline</td>
<td>Knowledge of research methods and conventions. Applies subject-specific knowledge to refine issue for investigation, identify question and conduct research. Own independent research using techniques and methods appropriate to the subject discipline i.e. literature search, relevant statistical/data handling and modelling techniques.</td>
</tr>
<tr>
<td>AO4a Communication in negotiating and conducting the research project</td>
<td>Communicate clearly Explanation and presentation of research methods, findings and conclusions</td>
<td></td>
</tr>
</tbody>
</table>

Note that AO4a was divided into AO4a and AO4b for the purposes of the research.

This study, which formed part of a wider programme of research into the IRR (Suto and Shaw, 2010; Shaw and Suto, 2010), investigated whether CI occurred when AO1 and AO4a were used to mark and internally moderate. It was conducted post hoc. A list of behaviours was taken from previous research, outside the Cambridge Pre-U context. The comments
made by the teacher and the IM on reports were searched for evidence of the behaviours. If several behaviours occurred, their relationship to marks was investigated. A systematic relationship between behaviours and marks would be indicative of CI.

Method

Data

92 candidates entered the IRR unit. The teacher and IM recorded comments on reports as appropriate. All available AO1 and AO4 comments were collected. There was a total of 150 comments: 67 comments about AO1 (60 from a teacher and seven from an IM) and 83 comments about AO4 (62 from a teacher and 21 from an IM). These comments were from a total of 70 candidates’ reports from eight centres.

Qualitative coding

Two researchers developed a qualitative coding system to analyses the comments. The coding system was developed from the behaviours noted in the literature (Table 2).

Table 2: Coding categories and associated behaviours found in the literature

<table>
<thead>
<tr>
<th>Behaviour noted in literature as a potential source of CI</th>
<th>Category description</th>
</tr>
</thead>
<tbody>
<tr>
<td>The assessor ……</td>
<td>The teacher/IM ……..</td>
</tr>
<tr>
<td>Compared a candidate’s performance with another candidate’s performance (Morgan, 1996; Crisp, 2010)</td>
<td>Compared a candidate’s performance with another candidate’s performance</td>
</tr>
<tr>
<td>Expresssed feelings towards a candidate e.g. hostility (Vaughan, 1991; Crisp, 2010)</td>
<td>Expresssed feelings towards a candidate</td>
</tr>
<tr>
<td>Laughed or noted amusement at a candidate or their performance (Vaughan, 1991; Crisp, 2010)</td>
<td>Expresssed amusement at a candidate’s performance/a candidate</td>
</tr>
<tr>
<td>Predicted the quality of a candidate’s future performance (Barritt et al., 1986; Crisp 2010)</td>
<td>Predicted the quality of a candidate’s future performance</td>
</tr>
<tr>
<td>Expressed a view on their own assessment practice (Crisp, 2010)</td>
<td>Expressed a view on their own summative assessment practice. NOT teaching/formative assessment.</td>
</tr>
<tr>
<td>Commented on a candidate’s characteristic such as skill/ability/gender (Barritt et al., 1986; Vaughan, 1991; Crisp 2010)</td>
<td>Commented on a candidate demographic/general ability</td>
</tr>
<tr>
<td>Used surface features of a candidate’s work in judgements (Morgan and Watson, 2002)</td>
<td>Referred to a surface feature(s) of a candidate’s work. NOT quality of written communication</td>
</tr>
<tr>
<td>Estimated a candidate’s effort invested in the work (Crisp, 2010)</td>
<td>Estimated a candidate’s effort invested in the work</td>
</tr>
</tbody>
</table>

Results

In the final analysis a total of five out of the 150 comments referred to a behaviour in the coding scheme. These behaviours occurred whilst assessing five out of 70 candidates.

- Two comments were categorised as “Expressed feelings towards a candidate”, they were:
  “CANDNAME was always very well informed, very well read, very focused and considered. Her desire to learn and complete this report was impressive from start to finish and she needed very little support from her supervisor”, was allocated to the category.
  “Communication highly effective both on paper and orally. Communicated in a mature and effective way with tutor. The candidate took considerable care to prepare for the meeting.” (Teacher’s comment regarding AO4)

- Three comments were categorised as “Referred to a surface feature(s) of candidate’s work in judgements”, they were:
  “Lack of intro and sustained argument – otherwise very good.” (Teacher’s comment regarding AO4)
  “Style is not academically formal in parts. Referencing not always clear or present. Clearly communicated well with tutor.” (IM’s comment regarding AO4)
  “Intro too long.” (IM’s comment regarding AO4)

Discussion

The research investigated whether CI occurred in the marking and internal moderation of AO1 and AO4a.

There were several limitations with the research. First, the list of behaviours was possibly inexhaustive. Second, the comments were a partial representation of each teacher’s and IM’s thoughts and deliberations. These limitations meant that some behaviours might be undetected. Third, the comments about AO4 did not differentiate between AO4a and AO4b. Therefore any behaviours related to AO4 cannot be attributed accurately to AO4a or AO4b. Despite these limitations the research evidence provides some useful and important findings.

Previous research, outside of the Pre-U context, identified several assessor behaviours which might be sources of CI if they systematically
influence marks. Two of these behaviours were noted in the comments. The teacher/IM “expressed feelings towards a candidate or a candidate’s performance” and “referred to a surface feature(s) of a candidate’s performance”. The behaviours occurred in comments about five out of 70 candidates (i.e. five out of 150 comments). These findings have resonance with previous findings about coursework/project work (Vaughan, 1991; Crisp, 2010; Morgan and Watson, 2002). The occurrence of these behaviours was not evidence of CI.

The lack of behaviours was a positive finding, particularly given that in several domains/professions erroneous information can influence judgement (Hackenbrack, 1992; Laming, 2004; Wistrich et al., 2005). Gaeth and Shanteau (1984) found that interactive training and practise reduced the influence of irrelevant information on the judgement of soil samples. Summers et al. (2004) found that formal education (rather than experience based learning) improved credit granting decisions. If experts in other domains can be trained to pay less attention to irrelevant information, then perhaps teachers and IMs can too. CIE runs standardisation meetings and provides other forms of centre support. This might well have contributed to the lack of behaviours found in the comments.

CI occurs only when such behaviours systematically influence marks. The lack of behaviours meant it was not possible to investigate a systematic relationship between the behaviours and marks. The result was that there was no evidence of CI.

There were different numbers of comments about AO1 (N=67) and AO4 (N=83). In other words some reports contained comments by the teacher/IM about AO1 but not AO4 and vice versa. This is not problematic as there was no requirement to make comments on reports, as noted earlier. However, it is interesting to consider why there were differences. The purposes of making annotations and comments on scripts might provide some insights. Crisp and Johnson (2007) and Fowles (2008) report two reasons for annotating and commenting whilst marking:

- Explaining decisions to others
- Supporting judgements and decision making during the process of marking

Perhaps there was a feeling that AO1 decisions were more self evident and needed fewer aide memoires than AO4 judgements.

Conclusion

This study found no evidence of CI in assessing the Knowledge and understanding of the research process (AO1) or Communication (AO4), and therefore no threats to validity were identified. This adds to the body of research supporting the teacher assessing the candidate’s IRR performance (Suto and Shaw, 2010; Shaw and Suto, 2010), the validity of the teacher assessment and internal moderation. The findings suggest that AO1 and AO4 facilitate valid assessment. Furthermore, standardisation and other forms of centre support may be useful in guarding against CI.

References


http://www.cie.org.uk/qualifications/academic/uppersec/preu/subjects/
subject/preusubject?assdef_id=1018


Bilingualism and bilingual education

Bilingualism is at least as common as monolingualism. Throughout the voluminous research literature, academic definitions of the term bilingual abound, from the early, limited and very narrow definitions, ‘native-like control of two languages’ (Bloomfield, 1935, p.56) to more flexible contemporary descriptions, ‘the presence of two or more languages’ (Dewaele, Houwen & Li, 2003, p.1), which reflect the awareness of the interdisciplinary nature of studies in bilingualism. Seen simply, the current view of bilingualism is the ability or need to perform in two languages.

Bilingual education—a simplistic label for a complex phenomenon (Baker, 2006, p.213)—has been widely researched since the mid-1960s, and experts throughout the world have attempted to define and analyse the complexity and effectiveness of different bilingual education programmes. Cummins (2009) offers a helpful conceptualisation of bilingual education: ‘an organized and planned program that uses two (or more) languages of instruction. The central defining feature of bilingual programs is that the languages are used to teach subject matter content rather than just the languages themselves.’ (p.161)

Interpretation of the research on bilingualism and bilingual education has been highly controversial among both academics and policy-makers, and political sensitivities surrounding the issue have contributed to considerable confusion about what the research is actually suggesting. However, over the last decade as knowledge of the extent of bilingualism has grown, discussions of bilingualism have focused on ‘the many kinds and degrees of bilingualism and bilingual situations’ (Crystal, 2003, p.51), leading to in-depth descriptions of the varied and disparate circumstances involved in bilingualism, anticipating the call for understanding the bilingual situation through its purpose and its context (Edwards, 2004).

The educational context

The context within which students prepare for Cambridge International Examinations (CIE) assessments are often linguistically and educationally diverse. Whatever the country, the common denominator of CIE schools is that students are taking CIE assessments through the medium of English and therefore being educated through the medium of English. Some schools use bilingual instruction, delivering certain subjects through English as an additional language and other subjects through the first language, often trying to meet standards in both an international curriculum and a national curriculum. The opportunity to learn an additional language through a content subject has led to the practice of content and language integrated learning (CLIL) programmes. Other schools use monolingual instruction and deliver all subjects through English, either as a first or as an additional language. Some of these schools will have monocultural student populations, whilst others will have multicultural populations. The latter places an additional responsibility on content teachers to be ‘language aware’ across the curriculum.

Whether students follow an entire curriculum in English or undertake only one or two CIE examinations in parallel with qualifications from their own (non-English) national curriculum as part of a bilingual education programme, the integration of curricula in bilingual education programmes presents challenges for teachers and their students. As a consequence, CIE is keen to understand this context in order to evaluate the impact of this choice of education programme and particularly the role of assessment within it. The Education Division (CIE) has outlined a coherent programme of research designed to address a number of key issues relating to bilingualism and learning, curriculum, pedagogy and assessment. The overall aim of the research is:

- to help build a pool of expertise and reputation in bilingual education
- to support the work of other parts of CIE involved in bilingual education, and
- to improve service to CIE schools.

The CIE bilingual research programme

The research programme is designed to address a number of specific questions grouped according to four principal themes:

1. **What is the impact of different teaching environments?**
   - What is the impact of different amounts of time studying in each language?
   - What is the impact of the choice of subject taught in the Second Language (L2)?
   - What is the impact of the way teachers share the language roles?
   - Are there significant differences in the way bilingual education is organised at primary and secondary level?

2. **What impact does bilingual education have on the teaching and learning process?**
   - Do academic skills and subject knowledge skills transfer across languages?
   - Do literacy skills transfer across languages?
   - What different approaches are taken to learning to read and write in bilingual education systems?
   - What research has been carried out about bilingual education in early years?
3. What is the impact of bilingual education on learner outcomes?
- What is the impact of bilingual education on achievement at school?
- What is the impact of having a bilingual education background for Higher Education?

4. What are the key assessment issues?
- What methods of assessment are used in bilingual education programmes?
- How is assessment adapted for bilingual education?

**Review of the literature**

Over the last 18 months, CIE has conducted a number of reviews of the bilingual education literature (Lewis, 2010; Chu et al., 2011) in order to begin to address some of these questions. The most salient points to emerge from the reviews are:

- More research is needed into ways of making academic content more accessible and meaningful to students in bilingual programmes particularly in areas/subjects considered challenging when learning academic content occurs through the L2.
- Bilingual education is a complex phenomenon.
- Different models of bilingual education – ‘weak’ (‘subtractive’) and ‘strong’ forms (‘additive’) – impact differently on learner outcomes and achievement at school.
- Traditional models over the last 40 years do not suffice in the twenty-first century.
- There has been a move away from traditional models of bilingual education and a focus more on effective classroom practice.
- Urgent need to develop effective bilingual assessment methods that reflect classroom practices.
- Effective evaluation of learning and understanding of emergent bilinguals through a framework of dynamic bilingualism and performance-based, on-going, multimodal assessments.
- A way to create more equitable assessments for emergent bilinguals is to employ ‘translanguaging’ practices within assessments. Baker (2000, p.104–105) defines translanguaging as ‘the hearing or reading of a lesson, a passage in a book or a section of work in one language and the development of the work (i.e. by discussion, writing a passage, completing a worksheet, conducting an experiment) in the other language’.

**What are the key messages?**

The key messages emerging from the literature reviews are that bilingual education is:

- **Challenging**, in terms of learning subjects through an additional language.
- **Complex**, with a discernible shift away from the simplicity and variety of typologies of bilingual education to ‘engaging with optimal classroom dual language practices that maximise growth and gains for individual children’ (Baker, 2008, p.106).
- **Changing**, and in need of urgent development, in terms of assessment. Traditional typologies are in need of expansion in order to capture the linguistic complexity of the emerging bilingual (and multilingual) education practices of the twenty-first century (Garcia, 2009a). In other words, there is a move away from effective models towards effective practices, and a shift from strict separation of languages to bilingual or multilingual discourse practices.

**Research questions**

CIE is therefore conducting a number of studies designed to address specific language-related questions in relation to its own assessments, as well as exploring the potential for bilingual assessments (formative or summative). For example:

- What level of English, according to the Common European Framework of References for Languages (CEFR), is useful to access CIE assessments?
- What cognitive and academic language skills are needed to access CIE assessments?
- How can schools be supported to prepare teachers and learners for whom English is a second language (L2) for bilingual programmes which include high-stakes, international assessment? For example, they could be given the tools to develop language awareness amongst content teachers as well as coordination between content and language teachers. Schools could also be guided on mapping their national against the international curriculum to streamline teaching and learning.
- Does the level of English impact on standards achieved in CIE non-language qualifications?
- What does successful attainment of CIE non-language qualifications indicate about language proficiency?
- What form of new assessments would enable bilingual students to demonstrate their strengths?

**Research into language awareness**

To tackle the first message of ‘challenge’ (and the first three research questions above), CIE has conducted research into the English language levels and skills required to achieve in typical CIE IGCSE subjects, focusing on History, Geography and Biology. By analysing assessment input and candidate output, CIE has identified useful target CEFR English language levels, as well as the type of cognitive academic language proficiency (CALP) needed for different subjects. For example:

<table>
<thead>
<tr>
<th>History</th>
<th>Biology</th>
<th>Geography</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainly open questions with constructed responses: describing, explaining, evaluating. In-depth source evaluation. E.g. <strong>How far...</strong></td>
<td>Precision: limited range of language functions signalled by different command words each with a precise meaning. E.g. <strong>Name...</strong></td>
<td>Flexibility: variety of question types requiring a range of language skills. Broad range of command words. E.g. <strong>Calculate...</strong> (prompting non-linguistic answers) <strong>Identify...</strong> (prompting short answers) <strong>Explain...</strong> (prompting developed answers)</td>
</tr>
<tr>
<td><em>Identify...</em>* (prompting short answers) <strong>Explain...</strong> (prompting developed answers)</td>
<td><em>Identify...</em>* (prompting short answers) <strong>Explain...</strong> (prompting developed answers)</td>
<td><em>Identify...</em>* (prompting short answers) <strong>Explain...</strong> (prompting developed answers)</td>
</tr>
</tbody>
</table>

... continues...
Practical applications have arisen from this research:

1) Language guidance has been provided for CIE question writers. For example:

- Avoid phrasal verbs: Phrasal verbs use simple words. However, they are colloquial and potentially difficult to process. Candidates may have difficulty with expressions like 'put up with'. This could be replaced by 'tolerate'.

2) Language guidance to support teachers and learners in the classroom has been commissioned by CIE (Chadwick, 2012). For example, this will encourage subject teachers to consider the following questions:

- What content vocabulary will my students need for the tasks in my lesson?
- How will I help my students with this vocabulary?
- What are students actually doing? What are the cognitive processes and creative thinking skills that they are using? What is the functional language that goes with these skills?
- How will I help my students with this language?
- What language skills are the students using? Do I want them to read, write, speak and/or listen?
- How will this affect the support I provide?

Research into bilingual education

To tackle the second message of ‘complexity’, CIE has commissioned a guide into excellence in bilingual education, capturing the key research on bilingual education as well as focusing on effective school management and classroom practice (Mehisto, 2012). This will focus on core considerations applicable to diverse contexts, regardless of the particular model of bilingual education that a school uses.

Research into developing bilingual assessments

Regarding the third message of ‘change’, CIE is at this stage simply exploring current thinking in order to consider future challenges and opportunities. Garcia (2009) notes that ‘no area of bilingual education is in more need of development than that of bilingual assessment.’ (p.378). She concludes that, ‘… without large-scale bilingual assessment that would take into consideration the bilingual continuum in which bilingual children operate, as well as the integrated nature of their language proficiency and content knowledge, bilingual children will never be able to demonstrate their strength’ (p.378).

Bilingual assessment is an issue that needs to be developed and researched in order to accommodate the bilingual continuum in which bilingual children operate. It is evident that this ‘most thorny issue’ (Garcia, 2009, p.396) stems from the fact that assessment methods for bilinguals have developed from the practice of testing monolinguals and most often ignore the children’s bilingualism by assessing their abilities and knowledge as if they were performing as two monolinguals. Consequently, the monolingualism of most assessments does not reflect the bilingual practices of the classroom (Garcia, 2009). There is a pressing requirement, therefore, to develop effective bilingual assessment methods that reflect classroom practices of using two (or more) languages for teaching and learning so that bilingual children are given the opportunity to show their proficiency and competences in both languages.

Garcia (2009, p. 371–375) has conducted a comprehensive review of ways in which all assessments, including large-scale standardised assessment, could be done in bilingual ways. One of her main recommendations includes a translanguage mode of bilingual assessment. Translanguaging ‘… is the act performed by bilinguals of accessing different linguistic features or various modes of what are described as autonomous languages, in order to maximize communicative potential’ (Garcia, 2009a, p.140). Translanguaging reinforces the interrelationship between the two languages while also reinforcing the languages. According to Baker (2000), translanguaging has the potential to ‘promote a deeper and fuller understanding of the subject matter’ (p.104) and to ‘develop skills in the weaker language …’ (p.105) and as a communicative practice, offers a range of communicative and educational possibilities.

The way forward

The growth of multiple multilingual education programmes at the end of the 20th century has been in response to the type of complex bilingualism brought about by globalisation (Garcia, 2009a, p.146). There is now a requirement to construct bilingual models that reflect the fluidity of classroom practices that have come about because of the ‘translanguaging’ that is characteristic of bilingual classrooms today. Echoing García’s concerns, it is crucial then that future bilingual assessment practice ‘can tap the pluriliteracies of multimodal texts which bilingual children must produce in the twenty-first century’ (Garcia, 2009, p.378).

Translanguaging in high-stakes situations may be only appropriate to localities where only two languages are in use (such as Welsh-English in Wales, or Spanish-English in certain parts of the USA). There would be considerable pragmatic issues in providing translanguaging opportunities on an equitable basis in multilingual contexts. Much of the literature focuses on the experience of classroom practice and formative assessment, and this may be where the potential lies for CIE schools.

CIE is attempting, therefore, to build an understanding of best practice in the area of bilingual education and to guide any future developments in terms of language awareness and bilingual assessment which can be shared more widely across the assessment community.

References


Cambridge Assessment Statistics Reports: Recent highlights

Joanne Emery, Tim Gill, Rebecca Grayson and Carmen L. Vidal Rodeiro Research Division

Introduction

The Research Division publishes a number of Statistics Reports each year based on the latest national examinations data. These are statistical summaries of various aspects of the English examination system, covering topics such as subject provision and uptake, popular subject combinations, trends over time in the uptake of particular subjects and the examination attainment of different groups of candidates.

The National Pupil Database (NPD) is the source of most of these reports. This is a very large longitudinal database, owned by the Department for Education, which tracks the examination attainment of all pupils within schools in England from their early years up to Key Stage 5 (A level or equivalent). It is updated annually from data provided by the awarding bodies and goes back as far as 1996. Another database, the Pupil Level Annual School Census (PLASC), can be requested matched to the NPD. This contains background information on candidates such as deprivation indicators, language, ethnicity and special educational needs.

Other sources of data used to produce the Statistics Reports include the Inter-Awarding Body Statistics produced by the Joint Council for Qualifications (JCQ) and the National Candidate Results Archive.

This article highlights some of the most recent Statistics Reports, published between 2010 and 2011. Full copies of all the Statistics Reports are available in the research section of the Cambridge Assessment website (www.cambridgeassessment.org.uk) and new additions to the Statistics Reports series will be listed in future issues of Research Matters.

Routine reports: Provision, uptake and results of GCSE and A level qualifications

A number of the statistics reports are produced routinely on a yearly basis. These reports are simple presentations of provision, uptake and results of GCSE and A levels, without commentary on the results. The purpose of these reports is to make readily available examinations data that is not (to our knowledge) provided elsewhere.

Uptake and results of GCSE and A level qualifications over time (Statistics Report Nos. 30–33)

The first set of routinely produced reports presents data on all entries and results for GCSEs and A levels taken in England, Northern Ireland and Wales over a period of several years (the latest reports are for 2002–2010). The data are compiled from the Inter-Awarding Body Statistics.

Four separate reports are routinely produced each year:
- GCSE uptake and results by gender
- A level uptake and results by gender
- GCSE uptake and results by school type
- A level uptake and results by school type

Within each report, uptake and results are presented for all subjects together and then broken down by subject category. Within each subject category there are sometimes different specifications. For instance, the
A level mathematics category includes Mathematics, Pure Mathematics, Further Mathematics and Statistics. Where individual subjects are of special interest these are presented separately. For instance, the modern languages category for both GCSE and A level was broken down into French, German and Spanish.

In each report the number of entries for each subject category is presented, followed by the cumulative percentage of candidates achieving each grade. Figure 1, taken from Statistics Report No. 31, presents the GCSE entries for all subjects for 2002–2010. An example of the trend in entries in an individual subject (GCSE Physics) is shown in Figure 2.

**Provision and uptake of GCSE and A level qualifications (Statistics Report Nos. 27–28)**

The second set of routinely produced reports is on the provision and uptake of GCSE and A level subjects in England in a given year, using data extracted from the NPD. Measuring these can identify subjects where levels of provision or uptake are low or declining.

The level of provision in a subject is defined as the percentage of schools with at least one student taking the subject. Uptake of a subject is measured as the percentage of all pupils taking at least one qualification of the same type (e.g. GCSE or A level) in the subject in question.

Four separate reports on provision and uptake are produced each year:
- Provision of A level subjects
- Uptake of A level subjects
- Provision of GCSE subjects
- Uptake of GCSE subjects

The levels of provision in the reports are presented by several school-level classifications: school type, school attainment, school gender composition (boys, girls or mixed), school size and school deprivation level. Similarly, uptake levels are reported by a number of student-level classifications: gender, school type, attainment, school gender composition and deprivation level.
Table 1: Provision of A levels by school type (percentages)

<table>
<thead>
<tr>
<th>Subject</th>
<th>Academy</th>
<th>Comprehensive</th>
<th>FE/Tertiary College</th>
<th>Grammar</th>
<th>Independent</th>
<th>Sec Mod</th>
<th>6th Form College</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td>73.0</td>
<td>92.0</td>
<td>72.9</td>
<td>99.4</td>
<td>87.1</td>
<td>56.4</td>
<td>94.8</td>
</tr>
<tr>
<td>Chemistry</td>
<td>66.1</td>
<td>88.4</td>
<td>68.6</td>
<td>99.4</td>
<td>84.4</td>
<td>42.6</td>
<td>96.3</td>
</tr>
<tr>
<td>English Literature</td>
<td>71.3</td>
<td>89.5</td>
<td>63.8</td>
<td>97.0</td>
<td>83.2</td>
<td>77.2</td>
<td>95.6</td>
</tr>
<tr>
<td>French</td>
<td>33.9</td>
<td>65.7</td>
<td>35.7</td>
<td>94.5</td>
<td>76.2</td>
<td>30.7</td>
<td>86.7</td>
</tr>
<tr>
<td>General Studies</td>
<td>16.5</td>
<td>37.0</td>
<td>8.7</td>
<td>59.8</td>
<td>15.8</td>
<td>14.9</td>
<td>42.2</td>
</tr>
<tr>
<td>Geography</td>
<td>47.8</td>
<td>83.0</td>
<td>44.4</td>
<td>98.2</td>
<td>80.1</td>
<td>57.4</td>
<td>88.1</td>
</tr>
<tr>
<td>History</td>
<td>67.8</td>
<td>90.6</td>
<td>65.7</td>
<td>99.4</td>
<td>83.5</td>
<td>64.4</td>
<td>94.1</td>
</tr>
<tr>
<td>Mathematics</td>
<td>79.1</td>
<td>94.4</td>
<td>73.4</td>
<td>99.4</td>
<td>90.4</td>
<td>73.3</td>
<td>97.8</td>
</tr>
<tr>
<td>Physics</td>
<td>55.7</td>
<td>83.7</td>
<td>60.9</td>
<td>99.4</td>
<td>82.3</td>
<td>38.6</td>
<td>94.8</td>
</tr>
<tr>
<td>Psychology</td>
<td>68.7</td>
<td>87.0</td>
<td>79.2</td>
<td>80.5</td>
<td>52.3</td>
<td>66.3</td>
<td>94.8</td>
</tr>
</tbody>
</table>

Table 2: Uptake of A levels by gender and prior attainment (percentages)

<table>
<thead>
<tr>
<th>Subject</th>
<th>All</th>
<th>Male</th>
<th>Female</th>
<th>Low attainment</th>
<th>Medium attainment</th>
<th>High attainment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics</td>
<td>24.4</td>
<td>31.5</td>
<td>18.4</td>
<td>6.8</td>
<td>19.2</td>
<td>44.5</td>
</tr>
<tr>
<td>Psychology</td>
<td>19.2</td>
<td>11.2</td>
<td>26.1</td>
<td>18.4</td>
<td>24.8</td>
<td>15.1</td>
</tr>
<tr>
<td>Biology</td>
<td>19.0</td>
<td>18.0</td>
<td>19.8</td>
<td>5.8</td>
<td>18.7</td>
<td>32.4</td>
</tr>
<tr>
<td>General Studies</td>
<td>16.9</td>
<td>17.0</td>
<td>16.8</td>
<td>11.1</td>
<td>18.4</td>
<td>22.3</td>
</tr>
<tr>
<td>History</td>
<td>16.3</td>
<td>17.4</td>
<td>15.4</td>
<td>10.7</td>
<td>16.6</td>
<td>22.3</td>
</tr>
<tr>
<td>English Literature</td>
<td>16.2</td>
<td>9.9</td>
<td>21.6</td>
<td>12.2</td>
<td>16.5</td>
<td>20.8</td>
</tr>
<tr>
<td>Chemistry</td>
<td>14.5</td>
<td>16.5</td>
<td>12.8</td>
<td>3.3</td>
<td>10.8</td>
<td>28.7</td>
</tr>
<tr>
<td>Geography</td>
<td>10.5</td>
<td>12.4</td>
<td>8.9</td>
<td>6.4</td>
<td>11.7</td>
<td>13.9</td>
</tr>
<tr>
<td>Business Studies: Single</td>
<td>10.5</td>
<td>13.5</td>
<td>7.9</td>
<td>13.2</td>
<td>12.9</td>
<td>5.2</td>
</tr>
<tr>
<td>Physics</td>
<td>10.1</td>
<td>17.4</td>
<td>3.9</td>
<td>2.7</td>
<td>8.5</td>
<td>18.1</td>
</tr>
</tbody>
</table>

Table 1 shows the provision of some of the most popular A level subjects in 2010 by school type (from Statistics Report No. 27).

Table 2 shows the levels of uptake of the ten most popular A level subjects in 2010 by gender and by prior attainment level (from Statistics Report No. 28). Prior attainment was defined by students’ GCSE grades. GCSE grades were converted into scores (A* = 8, A = 7, B = 6 etc.) and a mean GCSE was calculated for each student, which was then used to divide students up into three approximately equal-sized attainment groups: low, medium and high.

The uptake reports also present the most common combinations of three or more A level subjects taken by candidates. In 2010 the combination taken by the largest percentage of students was Biology, Chemistry and Mathematics, which was taken by 4% of male students and 4.2% of female students.

Provision and uptake of specific GCSE and A level subjects

At times, the provision or uptake of a particular subject area becomes a matter of public concern due, for example, to a rapid decline in pupil numbers or to the lack of choice within a particular school sector. On other occasions, changes in educational policy draw attention to specific subject areas. As a result, specific analyses on provision or uptake of GCSE or A level subject areas are carried out.

This section of the article outlines some of the results from two statistical reports investigating two subject areas, science and ICT, which have been in the spotlight recently. Data for these statistical reports were extracted from the NPD.

Provision of science subjects at GCSE 2009 (Statistics Report No. 15)

Recent changes to the level 2 curriculum have provided schools and students with a much greater choice of science qualifications. In particular, since 2006, the programme of study for science sets out a core content that is relevant to all pupils and specifies curriculum requirements for the equivalent of a single GCSE (Core Science). Additionally, students can take one of two complementary GCSEs – GCSE Additional Science or GCSE Additional Applied Science – in order to cover a more comprehensive programme of science study. Students can also study separate GCSE Biology, GCSE Chemistry and GCSE Physics to gain three full GCSEs in science. Since 2002 there has also been a vocational route in science offered at GCSE level: Applied Science Double Award. This qualification was designed to offer students the opportunity to widen their participation in vocationally-related learning.

Statistics Report No.15 investigated the provision of GCSE science options in secondary schools in England in 2009. The percentages of schools offering each science option were tabulated overall and by school type, school attainment and school deprivation. Table 3, showing the most popular combinations of science subjects offered in secondary schools in England both overall and by school type, highlights that:
- Biology, Chemistry and Physics were available for certification in 2009 in about 46% of the schools (increasing by about 11 percentage points since 2007 – see report);
- the provision of Biology, Chemistry and Physics was higher in independent and grammar schools than in comprehensive schools. The Double Award (Core plus Additional Science) and the Applied Science Double Award followed the opposite pattern.

Table 4: Entries for ICT and computing (or any related subjects), 2007–2009

<table>
<thead>
<tr>
<th>Level</th>
<th>Qualification</th>
<th>Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2007</td>
</tr>
<tr>
<td>1/2¹</td>
<td>GCSE full course in ICT</td>
<td>78414</td>
</tr>
<tr>
<td></td>
<td>GCSE short course in ICT/Digital Communications</td>
<td>77870</td>
</tr>
<tr>
<td></td>
<td>Vocational GCSE Double Award in Applied ICT</td>
<td>26470</td>
</tr>
<tr>
<td>2</td>
<td>Functional skills</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Key skills</td>
<td>6320</td>
</tr>
<tr>
<td></td>
<td>GNVQ in Applied ICT</td>
<td>48703</td>
</tr>
<tr>
<td></td>
<td>NVQs</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>VRQs</td>
<td>2310</td>
</tr>
<tr>
<td></td>
<td>Award/Certificate/Extended Certificate/Diploma in Digital Applications (DiDA)</td>
<td>68774</td>
</tr>
<tr>
<td></td>
<td>BTEC First for ICT practitioners</td>
<td>1393</td>
</tr>
<tr>
<td></td>
<td>OCR Nationals in ICT</td>
<td>5022</td>
</tr>
<tr>
<td></td>
<td>BCS²</td>
<td>5184</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>148</td>
</tr>
</tbody>
</table>

1 These qualifications are at levels 1 or 2 on the National Qualifications Framework, depending on the grade obtained. (http://www.direct.gov.uk/en/EducationAndLearning/QualificationsExplain/DG_10039017/).
2 BCS are qualifications awarded by The Chartered Institute for IT, formerly known as the British Computer Society.

There are other science qualifications at level 2 that account for a small percentage of the volume of science offered in schools (e.g. BTEC in Applied Science or OCR National Awards in Science). They are an alternative to the courses mentioned above and provide students with the technical knowledge and skills needed in the workplace, in further education or in training. The percentages of candidates taking these courses have been increasing over the last few years. Further work including these qualifications is currently ongoing.


The number of students taking ICT (Information and Communication Technology) and computing-related GCSE and A level qualifications has dropped in recent years, with a fall of 33% in just three years in ICT GCSE students, a fall of 33% in six years in A level ICT students and a fall of 57% in eight years in A level Computing students in England (The Royal Society, 2009).

Statistics Report No. 25 investigated trends in the numbers of students in England obtaining qualifications in ICT and Computing (or any related subjects) at Key Stage 4 and at Key Stage 5 over the years 2007 to 2009.

This report shows that in recent years many alternatives to GCSEs and A levels have been offered by the English awarding bodies (e.g. vocationally-related qualifications such as the Diploma in Digital Applications or the OCR Nationals). Some of these qualifications have become very popular among 14–19 year olds and some schools have moved away from GCSEs and A levels to take on vocational courses. In particular, Table 4 shows that:

- from 2007 to 2009, entries in GCSE ICT dropped both for the full course and for the short course (by 32% and 42%, respectively). There was also a fall of about 70% in the entries for the vocational GCSE in Applied ICT;
- the uptake of vocationally-related qualifications at level 2, such as BTEC Firsts, OCR Nationals and qualifications in the DiDA suite, experienced a large increase from 2007 to 2009.

Entries for level 3 qualifications, as well as entries by students’ ability, students’ level of deprivation and students’ school type, are also available in Statistics Report No. 25. The report also includes population estimates of 15 year olds to 18 year olds for the years 2007 to 2009 for England, which can be used to check for an increase or a decline in the population of students.

Other areas of research in recent Statistical Reports

How old are GCSE candidates? (Statistics Report No. 20)

Although GCSEs are designed for sixteen year olds, older and younger candidates can enter for them. For this report the distribution of GCSE entries and candidates by age is presented for three different years (2000, 2004 and 2009). The results are then broken down by what are considered to be important factors, such as school type and subject. Finally, the most popular subjects taken by candidates of different ages are shown. The data for this report come from the National Candidate Results Archive, which consists of all GCSE entries from all exam boards in England, Wales and Northern Ireland.

Table 5 presents all GCSE entries in each of the three years, broken down by candidate age.

As expected, the vast majority of entries were from pupils aged 16. The second highest number of entries were 17 year olds in 2000 and 2004 (making up 3.1% and 2.3% of entries), and 15 year olds in 2009 (5.1%). Indeed, 2009 saw a notable increase in the percentages of entries from 15 year olds in comparison to earlier years (from 1.7% in 2004 to 5.1% in 2009).

Tables presenting entries by age in a number of individual subjects are also included in this report. They show that, for example, there were substantial numbers of pupils aged 17 who were taking GCSE English or Maths, the majority of whom were likely to be re-taking the qualification. In 2009, both subjects had large increases in the percentages of pupils taking the qualification early compared with previous years. They were both also popular amongst adult learners,
presumably people obtaining a qualification to help them get into higher education or to get a job.

In French there were substantial percentages of early takers, mainly 15 year olds but many at age 14. This was also one of the most popular subjects for 11–13 year olds and many of these candidates may be native speakers of the language. There were large increases in the percentages of 14 and 15 year old takers in 2008 compared with earlier years.

Table 5: GCSE entries by age (all subjects)

<table>
<thead>
<tr>
<th>Age band</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 or younger</td>
<td>55</td>
<td>76</td>
</tr>
<tr>
<td>11–13</td>
<td>1,514</td>
<td>1,821</td>
</tr>
<tr>
<td>14</td>
<td>4,878</td>
<td>10,768</td>
</tr>
<tr>
<td>15</td>
<td>54,259</td>
<td>82,652</td>
</tr>
<tr>
<td>16</td>
<td>4,250,321</td>
<td>4,687,506</td>
</tr>
<tr>
<td>17</td>
<td>143,041</td>
<td>112,287</td>
</tr>
<tr>
<td>18</td>
<td>34,075</td>
<td>27,472</td>
</tr>
<tr>
<td>19</td>
<td>13,138</td>
<td>10,981</td>
</tr>
<tr>
<td>20–24</td>
<td>14,327</td>
<td>11,411</td>
</tr>
<tr>
<td>25–54</td>
<td>13,242</td>
<td>29,366</td>
</tr>
<tr>
<td>55 or older</td>
<td>3,372</td>
<td>2,895</td>
</tr>
</tbody>
</table>

Predicting A level grades using AS level grades (Statistics Report No. 29)

The university application process for candidates in the UK is run by UCAS. In the UCAS application process a referee for each candidate is required to submit predicted grades for the candidate’s pending qualifications. These referee-predicted grades are then used by universities and colleges to inform the offers made to their applicants.

The main qualification completed by candidates in England before university entry is the A level. A level qualifications are usually undertaken across a two year period, with candidates typically completing corresponding AS level qualifications at the end of the first year. For these candidates, a potential alternative to referee-predicted grades is therefore their actual AS level results.

The purpose of Statistics Report No. 29 was to explore the possibility of using AS level results as an alternative to predicted grades in the UCAS application process. Specifically, the report used the 2009 and 2010 NPD to analyse how accurately 2009 AS level grades were able to predict 2010 A level grades for candidates in England.

The report identified that 2009 AS level grades were a reasonable predictor of 2010 A level grades, with 54% of A level grades equal to AS level grades and 93% of A level grades within one grade of AS level grades (where the data could be matched). However, AS level grades predicted slightly disproportionately for some subgroups of candidates: they were more successful at predicting A level grades for candidates who attained high AS level grades, female candidates, candidates from areas of low or medium deprivation and candidates from independent or grammar schools.

A recent Department for Business, Innovation and Skills (BIS) report (2011) investigated how accurately predicted grades were able to predict A level grades for all UCAS applicants in the 2009 examination year. The results of this study allowed direct comparison of the predictive ability of AS level grades against that of referee-predicted grades, albeit within the limitations of different admission years. Table 6 presents the proportion of A level grades that were equal to, higher than, or lower than matched AS level versus predicted grade predictors in the two reports.

Table 6: A level grade equal to, higher than, or lower than predictor (AS level grade or predicted grade)

<table>
<thead>
<tr>
<th>A level grade</th>
<th>Predictor (column %)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AS level grade (2010 prediction)</td>
</tr>
<tr>
<td>Equal to predictor</td>
<td>54.5</td>
</tr>
<tr>
<td>Higher than predictor</td>
<td>22.9</td>
</tr>
<tr>
<td>Lower than predictor</td>
<td>22.7</td>
</tr>
</tbody>
</table>

Overall, AS level grades were a slightly more accurate predictor than predicted grades. Where the predictions were inaccurate, AS level grades were equally likely to be optimistic (A level lower than AS level) as they were pessimistic (A level higher than AS level). In contrast, predicted grades were substantially more likely to be optimistic than pessimistic. In the UCAS application process these varied balances of optimism relative to pessimism could have very different consequences for both the candidates and institutions involved (see Statistics Report No. 29 commentary for further discussion).

The BIS predicted grades report also investigated how accurately A level grades could be predicted for several different subgroups of candidates. Predicted grades were more successful at predicting A level grades for candidates who were predicted A grades, female candidates, candidates of high socio-economic status and candidates from independent schools. These patterns are very similar to those identified in the AS level grades report and highlight that AS level grades and predicted grades both predict slightly disproportionately for some subgroups of candidates.

Overall, the outcomes of the predicted grades report and Statistics Report No. 29 highlight that AS level grades could be considered as a possible alternative to (or supplement) referee-predicted grades. However, prediction of A level grades for both predictors could only be described as ‘reasonable’. The key question for consideration might therefore be whether either of the predictors is sufficiently accurate for use in the UCAS admission process.

Candidates awarded the new A* grade at A level in 2010 (Statistics Report No. 36)

Two previous Statistics Reports (No. 6 and No. 14) have shown the rise in the number and percentage of candidates, since 1996, attaining three or more A grades at A level. Of candidates in England aged 17–18, only around 8% of those taking at least three A levels attained three A grades or better in 1996 (less than 11,000 candidates). By 2006, this had risen to over 15% (more than 24,000 candidates). There followed a year-upon-year increase up to 2009, when the figure stood at around 17.5% of that group (around 30,000 candidates). Approximately 28% of all grades awarded in 2009 were an A. These attainment increases were problematic for a number of competitive higher education institutions and courses, which were faced with the task of differentiating between an increasing pool of equally highly-qualified applicants.
The A* grade, first awarded at A level in 2010, was designed to differentiate between the highest ability candidates. It is awarded to candidates who attain an A grade overall (at least 80% of the uniform marks across all their units) plus at least 90% of the uniform marks across their A2 (normally second-year) units. Statistics Report No. 36 investigated attainment of this new grade by candidates’ school type and gender.

In 2010, around 8% of A level entries were awarded the A* grade. Under 5000 A level pupils achieved three A* grades or better, this being less than 3% of the 180,181 17–18 year olds taking at least three A levels. Around 6% attained A*A*A or better and around 12% attained A*AA or better. The percentage attaining AAA or better decreased slightly from the 2009 figure, to just below 17%.

By school type, the percentage of pupils attaining three A* grades varied from around 6.5% in independent schools to under 1% in FE/tertiary colleges. It was around 1.5% in comprehensive schools and sixth form colleges. This is shown in Table 7. The results for grammar school pupils were closer to those of independent school pupils than they were to any other state school type. Over a third of independent school candidates attained grades AAA or better in 2010.

Table 8 shows that independent school candidates accounted for less than 16% of the A level candidature with three or more results but constituted almost 40% of the A*A*A* group. Comprehensive school candidates accounted for around 45% of the candiatde but only 25% of the A*A*A* group. Independent school candidates and comprehensive school candidates accounted for around 32% and 31%, respectively, of those achieving grades AAA or better. It can be seen from Table 8 that the greater the number of A* grades rather than A grades specified in the criterion, the more over-represented are candidates from independent and grammar schools.

The difference in A* grade attainment between males and females was much smaller than that between school types. Around 3% of males with three or more results, versus around 2% of females, achieved grades A*A*A* or better. Around 18% of males, versus 17% of females, achieved grades AAA or better. The A level candidature with three or more results is approximately 45% male. However, males became slightly more over-represented the higher the attainment criteria. Males comprised 54% of the candidates with grades A*A*A* or better and 47% of those with grades AAA or better.

References


---

Table 7: Percentages of candidates within each school type attaining the highest A level grades in 2010

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Comprehensive/Secondary Modern</th>
<th>Selective/Grammar</th>
<th>Independent</th>
<th>Sixth Form College</th>
<th>FE/Tertiary College</th>
</tr>
</thead>
<tbody>
<tr>
<td>A* A* A* or better</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>A* A* A or better</td>
<td>1.5</td>
<td>4.7</td>
<td>6.5</td>
<td>1.5</td>
<td>0.8</td>
</tr>
<tr>
<td>A* A A or better</td>
<td>3.9</td>
<td>10.6</td>
<td>14.3</td>
<td>4.2</td>
<td>2.4</td>
</tr>
<tr>
<td>A A A or better</td>
<td>7.9</td>
<td>19.2</td>
<td>25.3</td>
<td>8.4</td>
<td>5.2</td>
</tr>
<tr>
<td>A A A or better</td>
<td>11.4</td>
<td>26.9</td>
<td>35.0</td>
<td>12.3</td>
<td>8.0</td>
</tr>
</tbody>
</table>

Table 8: School type breakdown of the candidates attaining the highest A level grades in 2010

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Comprehensive/Secondary Modern</th>
<th>Selective/Grammar</th>
<th>Independent</th>
<th>Sixth Form College</th>
<th>FE/Tertiary College</th>
<th>Total N of candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>A* A* A* or better</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>Count</td>
</tr>
<tr>
<td>A* A* A or better</td>
<td>25.5</td>
<td>20.8</td>
<td>39.4</td>
<td>11.7</td>
<td>2.5</td>
<td>4639</td>
</tr>
<tr>
<td>A* A A or better</td>
<td>28.2</td>
<td>19.2</td>
<td>36.0</td>
<td>13.4</td>
<td>3.0</td>
<td>11180</td>
</tr>
<tr>
<td>A A A or better</td>
<td>30.0</td>
<td>18.4</td>
<td>33.7</td>
<td>14.4</td>
<td>3.4</td>
<td>21232</td>
</tr>
<tr>
<td>A A A or better</td>
<td>30.5</td>
<td>18.2</td>
<td>32.8</td>
<td>14.7</td>
<td>3.7</td>
<td>30144</td>
</tr>
<tr>
<td>Candidates with 3+ results</td>
<td>44.9</td>
<td>11.3</td>
<td>15.7</td>
<td>20.1</td>
<td>7.8</td>
<td>180181</td>
</tr>
</tbody>
</table>

Data source: National Pupil Database (DfE). Results are for pupils in schools in England who turned 18 in the school year 2009/10. General Studies and Critical Thinking results are excluded as some HE courses will not accept these other than as a fourth subject.
Conferences and seminars

Society for Research in Higher Education (SHRE)
In December 2011, Irenka Suto attended the SHRE annual conference in Cardiff, Wales. The theme of the conference was ‘Positive futures for higher education: connections, communities and criticality’. The keynote speakers were Professor Hugh Lauder, Professor of Education and Political Economy at the University of Bath, and Professor Maria Helena Nazaré, President of the European Universities Association.

NPD/PLASC Users Group
Rebecca Grayson and Tim Gill attended the NPD/PLASC Users Group conference in Bristol in March and presented a paper on ‘The use of the NPD and PLASC in examinations research’.

Sixth annual Rasch User Group meeting
In March Tom Bramley attended the sixth UK Rasch User Group meeting in Leeds. The Group provides a forum for Rasch enthusiasts working in different fields to get together to share ideas and present research. The purpose of the Group is to offer advice, support and encouragement to anyone interested in the Rasch model.

National Council on Measurement in Education (NCME)
The NCME annual meeting took place in Vancouver, Canada, in April and was attended by Tom Bramley.

UCAS admissions conference
In April Irenka Suto presented a paper at the UCAS Admissions Conference in Birmingham entitled: ‘What is the role of the A level in ensuring a successful transition from level 3 study to university education?’. The conference featured a range of high profile speakers who focussed on the challenges, changes and wider issues affecting admissions to higher education.

British Psychological Society Annual Conference (BPS)
Sanjana Mehta attended the BPS annual conference in April and presented a paper on ‘Is Psychology A level fit for purpose? Views from Higher Education’. The themes of the conference for 2012 were:
- The psychology of participation in sports and exercise
- The psychology of violence and conflict
- Language and communication
- Psychology for the public and private sectors

Publications

The following articles have been published since issue 13 of Research Matters:


CONTENTS : Issue 14 June 2012

2  An intra-board comparison of the effect of using pseudo candidates' scripts and real candidates' scripts in a rank-ordering exercise at syllabus level : Louis Yim

10 The effect of scripts' profiles upon comparability judgements : Nicky Rushton

18 Monitoring the difficulty of tiered GCSE components using threshold marks for grade C : Vikas Dhawan

21 An investigation on the impact of GCSE modularisation on A level uptake and performance : Carmen L. Vidal Rodeiro

29 Piloting a method for comparing the demand of vocational qualifications with general qualifications : Jackie Greatorex and Hannah Shiell

38 The validity of teacher assessed Independent Research Reports contributing to Cambridge Pre-U Global Perspectives and Research : Jackie Greatorex and Stuart Shaw

42 The Cambridge International Examinations bilingual research agenda : Helen Imam and Stuart Shaw

45 Cambridge Assessment Statistics Reports: Recent highlights : Joanne Emery, Tim Gill, Rebecca Grayson and Carmen L. Vidal Rodeiro

51 Research News

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

Tel: 01223 553854
Fax: 01223 552700
Email: ResearchProgrammes@cambridgeassessment.org.uk
http://www.cambridgeassessment.org.uk

© UCLES 2012

ISSN: 1755–6031