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" (p1), 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

1. Comparability studies are comparisons of qualifications, either investigating the same qualification and subject over different examination sessions, or making comparisons across different subjects or related qualifications. These comparisons can be made using statistical methods, by comparing the content of qualification, or by comparing examples of performance, e.g. scripts. The latter encompasses methods involving holistic judgements.

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[2] 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.

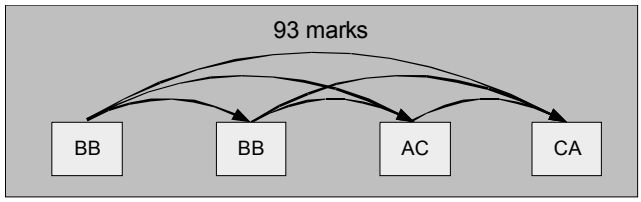

Figure 1: Comparisons of scripts at a single mark point

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.

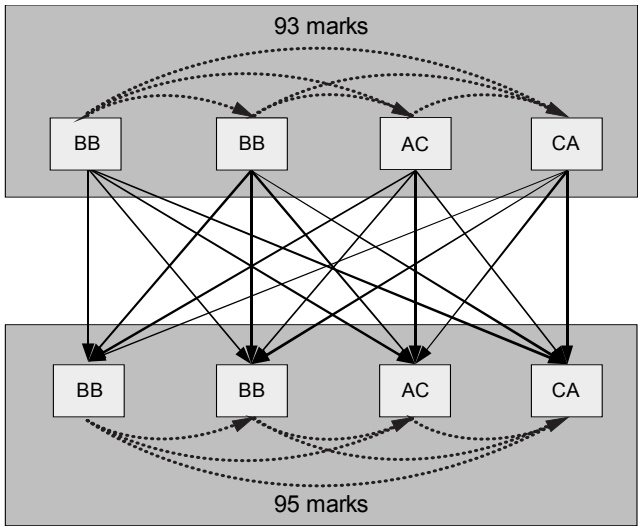

Figure 2: Comparisons of scripts with two different marks

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

	Judge 1	*Judge 2*	*Judge 3*	*Judge 4*	*Judge 5*	*All*
Even profile wins	27	29	33	27	26	142
Even profile loses	33	31	27	33	34	158

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

	Judge 1	*Judge 2*	*Judge 3*	*Judge 4*	*Judge 5*	*All*
Even profile wins	17	17	21	16	17	88
Even profile loses	19	19	15	20	19	92

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

	Judge 1	*Judge 2*	*Judge 3*	*Judge 4*	*Judge 5*	*All*
Higher score wins	8	8	8	9	9	42
Lower score wins	16	16	16	15	15	78

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

Winning script	*Judge 1*	*Judge 2*	*Judge 3*	*Judge 4*	*Judge 5*	*All*
Higher score wins						
Even profile wins with higher score	3	4	4	4	3	18
Uneven profile wins with higher score	5	4	4	5	6	24
Lower score wins						
Even profile wins with lower score	7	8	8	7	6	36
Uneven profile wins with lower score	9	8	8	8	9	42

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 5: Comparison by profile of script

	Judge 1	*Judge 2*	*Judge 3*	*Judge 4*	*Judge 5*	*All*
Even vs. better F324 result						
Even profile wins	15	16	15	13	13	72
Better F324 wins	15	14	15	17	17	78
Even vs. better F325 result						
Even profile wins	12	13	18	14	13	70
Better F325 wins	18	17	12	16	17	80
Better F324 vs. better F325 result						
Better F324 wins	7	6	9	7	8	37
Better F325 wins	8	9	6	8	7	38

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

	Judge 1	*Judge 2*	*Judge 3*	*Judge 4*	*Judge 5*	*All*[3]
Even profile wins	17	24	24	21	28	114
Even profile loses	39	32	32	35	27	165

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

	Judge A	*Judge B*	*Judge C*	*Judge D*	*Judge E*	*All*
Even profile wins	10	13	14	14	16	67
Even profile loses	24	21	20	20	17	102

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

	Judge A	*Judge B*	*Judge C*	*Judge D*	*Judge E*	*All*
Higher score wins	13	13	12	9	8	55
Lower score wins	9	9	10	13	14	55

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 56 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

	Judge A	*Judge B*	*Judge C*	*Judge D*	*Judge E*	*All*
Higher score wins						
Even profile wins with higher score	4	6	5	2	4	21
Uneven profile wins with higher score	9	7	7	7	4	34
Lower score wins						
Even profile wins with lower score	3	5	5	5	8	26
Uneven profile wins with lower score	6	4	5	8	6	29

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 AS unit than they did on the A2 unit (better AS 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

	Judge A	*Judge B*	*Judge C*	*Judge D*	*Judge E*	*All*
Even vs. better AS result						
Even profile wins	7	11	9	7	14	48
Even profile loses	19	15	17	19	12	82
Even vs. better A2 result						
Even profile wins	10	13	15	14	14	66
Even profile loses	20	17	15	16	15	83
Better AS result vs. better A2 result						
Better AS result wins	8	7	7	11	7	40
Better A2 result wins	5	6	6	2	5	24

For the candidates with a better A2 result, the judges' decisions were more inconclusive. Judges A and B appeared to favour the uneven profiled 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

	Judge 1	*Judge 2*	*Judge 3*	*Judge 4*	*Judge 5*	*All*
Even profile vs. better F324 result	3.1	3.2	3.7	3.3	2.9	3.3
Even profile vs. better F325 result	3.0	3.3	3.3	3.3	2.7	3.1
Even profile vs. even profile	2.6	3.0	3.4	2.8	2.1	2.8
Better F324 result vs. better F325 result	3.7	3.4	3.3	3.1	3.6	3.4
Both scripts better F324 result	5.0	3.7	4.0	4.7	3.0	4.1
Both scripts better F325 result	4.3	3.7	2.7	1.7	0.7	2.6

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 judges' ratings for every combination of profiles involved in the comparison. These were then totalled across particular types of comparison.

Table 12: Difficulty ratings of English Literature judges for Comparison 1 (comparing pairs of units)

	Judge A	*Judge B*	*Judge C*	*Judge D*	*Judge E*	*All*
Even profile vs. better AS result	3.0	2.7	2.8	2.9	2.4	2.8
Even profile vs. better A2 result	3.3	2.8	2.8	2.9	1.8	2.7
Even profile vs. even profile	3.6	2.2	2.9	3.2	1.4	2.7
Better AS result vs. better A2 result	3.2	2.8	2.3	2.8	2.0	2.6
Both scripts better AS result	2.5	1.5	2.0	3.0	2.0	2.2
Both scripts better A2 result	3.3	2.3	3.0	2.3	2.7	2.7

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 profile scripts had a higher score than the uneven profile 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 profile 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

Bramley, T. (2007). Paired comparison methods. In: P. Newton, J.-A. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Bramley, T. (2012). The effect of manipulating features of examinees' scripts on their perceived quality. *Research Matters: A Cambridge Assessment Publication*, **13**, 18–26.

Cresswell, M. (1997). *Examining judgements: Theory and practice of awarding public examination grades*. PhD thesis, Institute of Education, University of London.

Crisp, V. (2010). Judging the grade: exploring the judgement processes involved in examination grading decisions. *Evaluation & Research in Education*, **23**, 1, 19–35.

Edwards, E. & Adams, R. (2002). *A comparability study in GCE Advanced Level Geography including the Scottish Advanced Higher Grade examinations. A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2001 Examination*. Organised by WJEC on behalf of the Joint Council for General Qualifications, London.

Elliott, G. & Greatorex, J. (2002). A fair comparison? The evolution of methods of comparability in national assessment. *Educational Studies*, **28**, 3, 253–264.

Pollitt, A. & Elliott, G. (2003). *Monitoring and investigating comparability: a proper role for human judgement*. Cambridge: Research and Evaluation Division, University of Cambridge Local Examinations Syndicate.

Scharaschkin, A. (1997). *What do 'balanced candidates' tell us about standards? Internal report RAC/766*. Guildford: Associated Examining Board.

Scharaschkin, A. & Baird, J.-A. (2000). The effects of consistency of performance on A level examiners' judgements of standards. *British Educational Research Journal*, **26**, 3, 343–357.

Thurstone, L.L. (1927). A law of comparative judgement. *Psychological Review*, **34**, 273–286.

Yim, L. W. K. & Forster, M. (2010). A comparison between the effect of using pseudo-candidates' scripts and real-candidate's scripts in a rank-ordering comparability methodology at syllabus level. *36th IAEA Conference Proceedings – Assessment for the future generations*, Bangkok, 22–27 August 2010.