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[1], 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[2] within the same subject (Yim, Shaw and Lewis, 2008; Yim and Shaw, 2009).

The notion of a pseudo (or composite) candidate has been adopted for syllabus level comparability exercises. A pseudo 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 (Arlett, 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[3] 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[4] 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[5]; 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 ±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

Types of scripts	$\Delta_{measure}$ *[logit]*				
	A	*B*	*C*	*D*	*E*
Pseudo candidates	0.49	0.33	0.30	0.21	0.24
Real candidates	0.27	0.37	0.61	0.66	0.86

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 $\Delta_{measure}$, 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

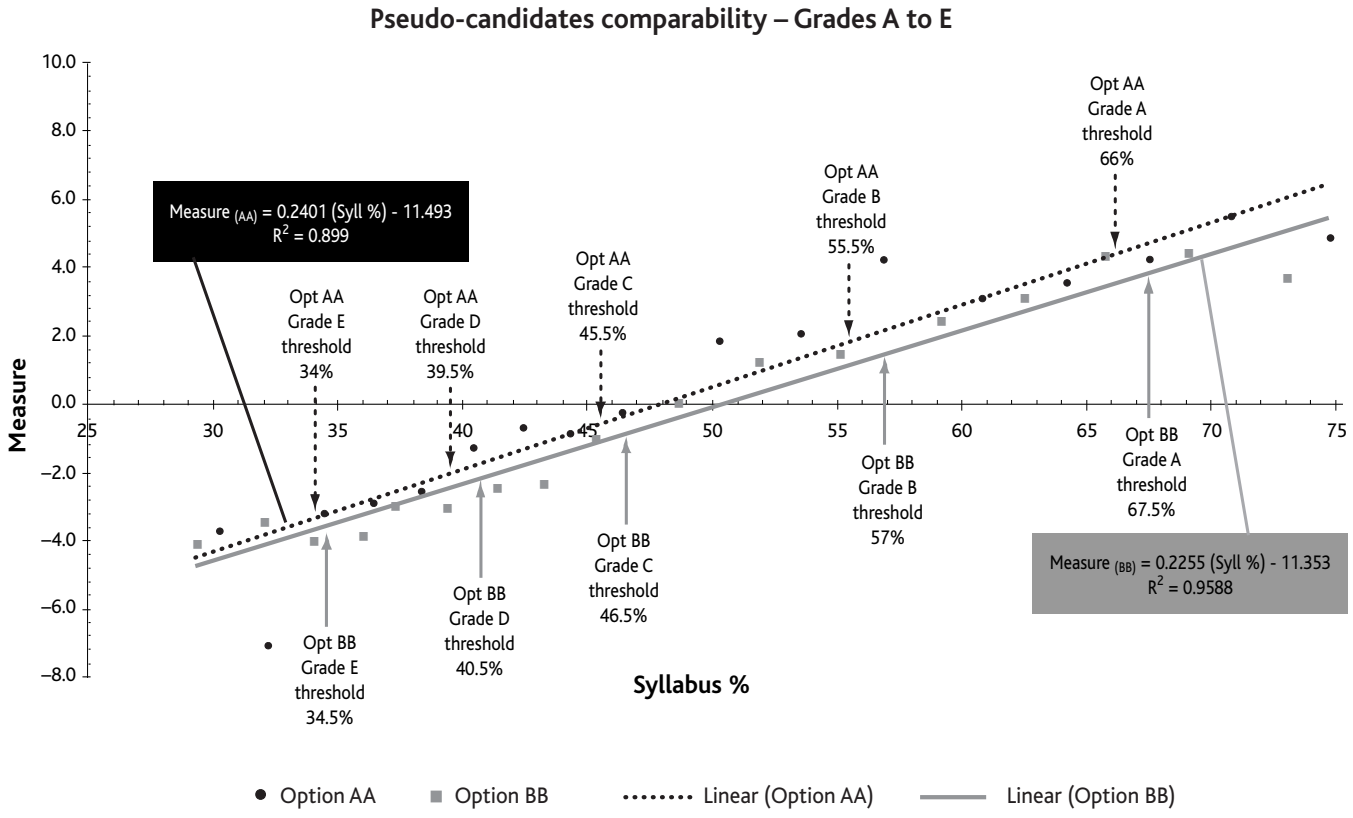

Figure 1: A comparability plot from grades A to E for pseudo candidates' scripts between Options AA and BB

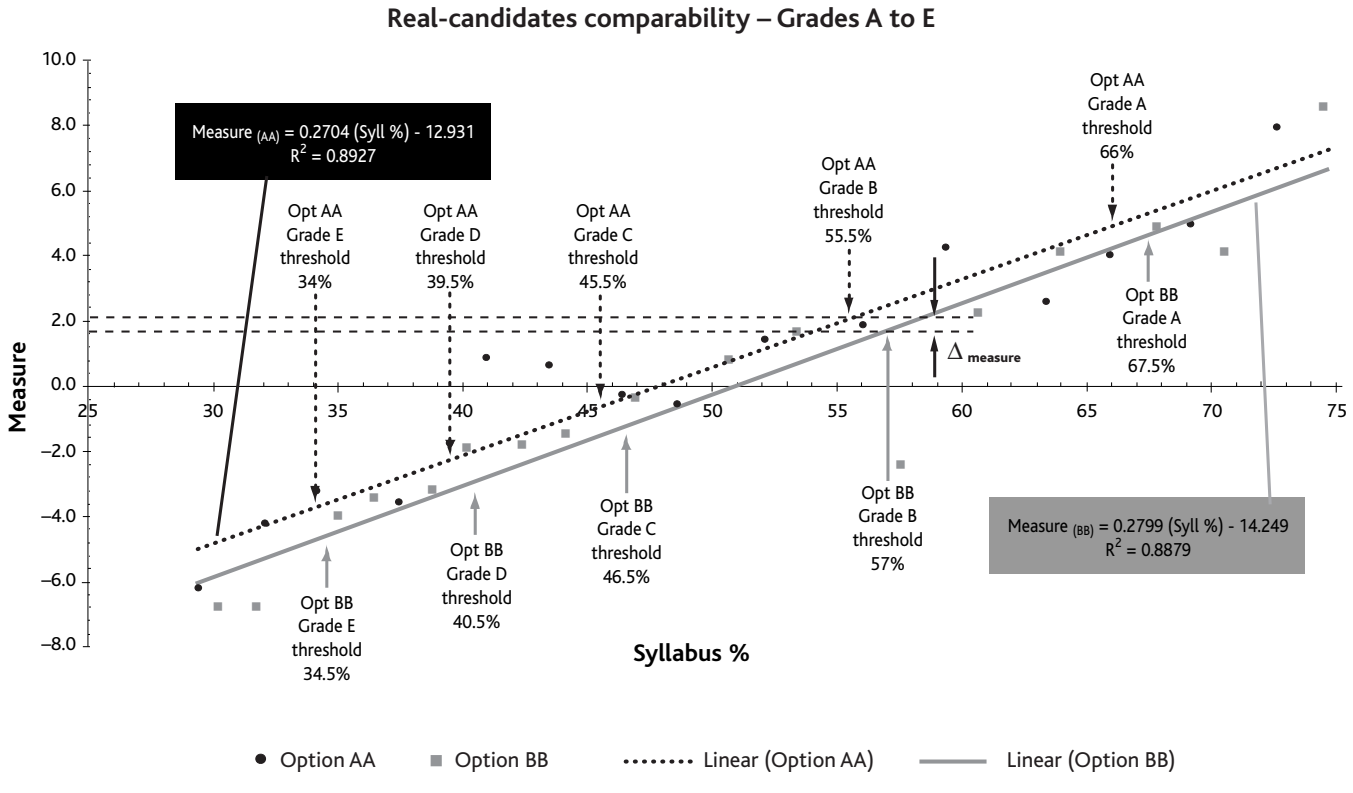

Figure 2: A comparability plot from grades A to E for real candidates' scripts between Options AA and BB

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

Assessment	*Type of scripts*	*Correlation coefficient (R)*
AA	Pseudo candidates	0.948
	Real candidates	0.945
BB	Pseudo candidates	0.979
	Real candidates	0.942

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.

Table 3: Difficulties encountered by examiners during the two evaluation phases

Pseudo candidates' scripts	*Real candidates' scripts*
• differences between questions in question papers from both options; • difficult to retain script information long enough to make judgement on the rank-order; • inconsistent quality/style across pseudo candidates' profile; • pseudo candidates' standards are very close within each pack.	• differences between questions in question papers from both options; • difficult to obtain an overview of papers with a number of parts; • difficult to retain script information to make judgement on the rank-order; • real candidates' standards are very close within each 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.

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgement. *Journal of Applied Measurement*, **6**, 2, 202–223.

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

Guthrie, K. (2003). A comparability study in GCE Business Studies, Units 4, 5 and 6 VCE Business, units 4, 5 and 6. A review of the examination requirements and a report on the cross-moderation exercise. A study based on the summer 2002 examination. Organised by Edexcel on behalf of the Joint Council for General Qualifications.

Hambleton, R.K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In: G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods and Perspectives*. 89–116. Mahwah, NJ: Lawrence Erlbaum Associates.

Jones, B., Meadows, M. & Al-Bayatti, M. (2004). Report of the inter-awarding body comparability study of GCSE religious studies (full course) summer 2003. Assessment and Qualifications Alliance.

Linacre, J.M. (2008). FACETS Rasch measurement computer program. Chacago: Winsteps.com.

Newton, P. (2007). Contextualising the comparability of examination standards. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms. (Eds.), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Rushton, N. (2012). The effect of scripts' profiles upon comparability judgements. *Research Matters: A Cambridge Assessment Publication*, **14**, 10–17.

Thurstone, L.L. (1927). A law of comparative judgement. *Psychological Review*, **34**, 273–286. Chapter 3 in L.L. Thurstone (1959), *The measurement of values*. Chicago, Illinois: University of Chicago Press.

Yim, L.W.K., Shaw, S.D. & Lewis, M. (2008). *A science comparability study between two exam boards using a rank-ordering methodology at syllabus level*. Paper presented at 9th AEA Europe Conference Proceeding, Hisar, Bulgaria, 6–8 Nov 2008.

Yim, L.W.K & Shaw, D. S. (2009). *A comparability study using a rank-ordering methodology at syllabus level between examination boards*. Paper presented at 35th IAEA Annual Conference Proceedings, Brisbane, Australia, 13–18 September 2009.

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

Appendix A: FACETS output

i) Pseudo candidate scripts' output

Table 7.1.1 Judge Measurement Report (arranged by mN)

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N Judge
60	120	.5	.50	.00	.21	1.20	2.0	1.20	1.1	.59	.39	.51	1 CHS
60	120	.5	.50	.00	.21	1.01	.1	1.14	.8	.95	.49	.51	2 TC
60	120	.5	.50	.00	.21	1.01	.1	1.01	.1	.97	.50	.51	3 PC
60	120	.5	.50	.00	.21	.79	-2.5	.69	-2.0	1.47	.64	.51	4 NB
60	120	.5	.50	.00	.21	.95	-.5	.89	-.6	1.12	.54	.51	5 GM
60.0	120.0	.5	.50	.00	.21	.99	-.1	.98	-.1		.51		Mean (Count: 5)
.0	.0	.0	.00	.00	.00	.13	1.5	.18	1.2		.08		S.D. (Population)
.0	.0	.0	.00	.00	.00	.15	1.6	.20	1.3		.09		S.D. (Sample)

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 7.3.1 Script Measurement Report (arranged by mN)

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu Script
12.5	25	.5	1.00	5.49	.47	1.04	.2	1.07	.3	.92	.50	.53	18 t1_P_A1 (mark 70.8)
12.5	25	.5	.99	4.86	.43	.91	-.5	.91	-.5	1.38	.45	.34	20 t1_P_A5 (mark 74.8)
12.5	25	.5	.99	4.42	.42	.99	.0	1.00	.0	1.06	.31	.29	1 t2_P_A2 (mark 69.3)
25	50	.5	.99	4.37	.31	1.05	.4	1.01	.1	.89	.39	.42	3 t2_P_A6B4 (mark 66)
12.5	25	.5	.99	4.22	.63	.91	.0	.69	-.3	1.10	.80	.77	23 t1_P_C4 (mark 56.8)
25	50	.5	.99	4.21	.31	.94	-.5	.94	-.3	1.20	.46	.40	19 t1_P_A4B3 (mark 67.5)
12.5	25	.5	.98	3.67	.45	1.01	.1	1.04	.2	.96	.45	.47	2 t2_P_A3 (mark 73.3)
12.5	25	.5	.97	3.52	.43	1.12	.8	1.14	.9	.48	.19	.35	21 t1_P_B1 (mark 64.2)
25	50	.5	.96	3.11	.32	.91	-.5	.91	-.4	1.19	.54	.48	22 t1_P_B6C5 (mark 60.8)
12.5	25	.5	.96	3.08	.43	1.15	1.0	1.17	.9	.46	.21	.38	5 t2_P_B5 (mark 62.7)
25	50	.5	.92	2.44	.33	1.01	.1	.95	-.2	1.02	.52	.52	4 t2_P_B2C2 (mark 59.3)
25	50	.5	.89	2.09	.35	.98	.0	.92	-.1	1.04	.60	.58	24 t1_P_C6D6 (mark 53.5)
12.5	25	.5	.86	1.84	.52	1.02	.1	.80	-.3	1.04	.65	.64	25 t1_P_D2 (mark 50.2)
12.5	25	.5	.81	1.44	.48	.90	-.4	.72	-.6	1.24	.62	.55	6 t2_P_C1 (mark 55.3)
25	50	.5	.76	1.17	.35	1.02	.1	.93	-.1	.99	.58	.58	7 t2_P_C3D3 (mark 52)
12.5	25	.5	.50	.02	.49	1.10	.4	1.18	.6	.82	.51	.57	8 t2_P_D1 (mark 48.7)
25	50	.5	.43	-.27	.35	1.07	.4	1.06	.3	.89	.54	.57	26 t1_P_D5E3 (mark 46.3)
25	50	.5	.33	-.69	.35	1.14	.8	1.22	.8	.77	.52	.59	27 t1_P_E2F3 (mark 42.3)
12.5	25	.5	.30	-.86	.44	.93	-.3	.90	-.4	1.23	.50	.43	28 t1_P_E6 (mark 44.2)
25	50	.5	.27	-1.01	.35	.68	-2.1	.52	-1.7	1.55	.74	.58	9 t2_P_D4E1 (mark 45.4)
12.5	25	.5	.22	-1.29	.49	1.07	.3	1.09	.4	.90	.54	.58	29 t1_P_F1 (mark 40.3)
12.5	25	.5	.09	-2.35	.50	.99	.0	1.07	.3	.99	.60	.60	10 t2_P_E4 (mark 43.3)
25	50	.5	.08	-2.49	.34	.92	-.4	.95	-.1	1.12	.58	.54	11 t2_P_E5F6 (mark 41.4)
25	50	.5	.07	-2.52	.31	.87	-1.0	.82	-1.1	1.38	.56	.44	30 t1_P_F4G5 (mark 38.2)
12.5	25	.5	.05	-2.89	.42	1.03	.2	1.04	.2	.85	.29	.33	31 t1_P_G2 (mark 36.3)
25	50	.5	.05	-2.99	.31	1.20	1.7	1.39	2.1	.23	.21	.41	12 t2_P_F2G3 (mark 37.3)
12.5	25	.5	.04	-3.10	.47	.94	-.2	.87	-.3	1.14	.56	.52	13 t2_P_F5 (mark 39.4)
25	50	.5	.04	-3.19	.31	.93	-.7	.84	-.4	1.30	.46	.41	32 t1_P_G6H6 (mark 34.3)
12.5	25	.5	.03	-3.53	.45	1.13	1.0	2.01	1.7	.07	.33	.46	16 t2_P_H3 (mark 32)
12.5	25	.5	.02	-3.69	.45	.97	-.2	.81	-.3	1.21	.47	.44	34 t1_P_H5 (mark 30.1)
12.5	25	.5	.02	-3.91	.44	1.08	.4	1.08	.4	.78	.35	.42	15 t2_P_G4 (mark 36)
25	50	.5	.02	-4.04	.32	.92	-.6	.83	-.7	1.25	.52	.46	14 t2_P_G1H1 (mark 34)
12.5	25	.5	.02	-4.18	.46	.93	-.3	.80	-.4	1.25	.53	.47	17 t2_P_H4 (mark 29.3)
12.5	25	.5	.00	-6.96	1.02	1.02	.3	1.27	.5	.97	.91	.92	33 t1_P_H2 (mark 32.2)
17.6	35.3	.5	.49	.00	.43	1.00	.0	1.00	.0		.50		Mean (Count: 34)
6.2	12.3	.0	.42	3.31	.13	.10	.7	.25	.7		.15		S.D. (Population)
6.2	12.5	.0	.42	3.36	.13	.10	.7	.25	.8		.16		S.D. (Sample)

Model, Populn: RMSE .45 Adj (True) S.D. 3.28 Separation 7.36 Reliability .98
Model, Sample: RMSE .45 Adj (True) S.D. 3.33 Separation 7.47 Reliability .98
Model, Fixed (all same) chi-square: 2148.4 d.f.: 33 significance (probability): .00
Model, Random (normal) chi-square: 32.3 d.f.: 32 significance (probability): .45

ii) Real candidate scripts' output

Table 7.1.1 Judge Measurement Report (arranged by mN)

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N Judge
60.5	121	.5	.50	.00	.24	1.02	.1	.89	-.3	1.00	.61	.61	3 PC
60	120	.5	.50	.00	.24	1.15	1.3	1.00	.0	.80	.57	.61	1 CHS
60	120	.5	.50	.00	.24	.88	-1.0	.70	-1.1	1.22	.66	.61	2 TC
60	120	.5	.50	.00	.24	1.20	1.7	1.37	1.3	.63	.53	.61	4 NB
60	120	.5	.50	.00	.24	.76	-2.3	.54	-2.0	1.44	.71	.61	5 GM
60.1	120.2	.5	.50	.00	.24	1.00	.0	.90	-.4		.62		Mean (Count: 5)
.2	.4	.0	.00	.00	.00	.17	1.5	.28	1.1		.06		S.D. (Population)
.2	.4	.0	.00	.00	.00	.18	1.7	.32	1.3		.07		S.D. (Sample)

Model, Populn: RMSE .24 Adj (True) S.D. .00 Separation .00 Reliability 1.00
Model, Sample: RMSE .24 Adj (True) S.D. .00 Separation .00 Reliability .80
Model, Fixed (all same) chi-square: .0 d.f.: 4 significance (probability): 1.00

Table 7.3.1 Script Measurement Report (arranged by mN)

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu Script
12.5	25	.5	1.00	8.65	.83	.85	-.1	.23	-.3	1.19	.89	.87	18 t2_R_A3 (mark 74.8)
12.5	25	.5	1.00	7.95	.75	1.22	.5	1.03	.3	.87	.82	.84	2 t1_R_A2 (mark 72.6)
12.5	25	.5	.99	4.97	.52	1.26	1.1	1.48	.8	.49	.53	.63	1 t1_R_A1 (mark 69.2)
25	50	.5	.99	4.90	.36	.87	-.8	.62	-.9	1.28	.67	.61	19 t2_R_A4B2 (mark 68.1)
25	50	.5	.99	4.25	.40	.99	.0	.84	-.3	1.05	.65	.64	4 t1_R_B4C3 (mark 59.3)
12.5	25	.5	.98	4.17	.53	.94	-.2	.60	-.1	1.20	.67	.64	20 t2_R_A6 (mark 70.8)
12.5	25	.5	.98	4.14	.45	1.08	.5	1.12	.5	.78	.39	.46	22 t2_R_B3 (mark 64.2)
25	50	.5	.98	4.06	.34	1.04	.3	.86	.0	.94	.54	.55	3 t1_R_A5B5 (mark 65.9)
12.5	25	.5	.93	2.61	.51	.96	.0	.88	-.2	1.07	.65	.62	5 t1_R_B6 (mark 63.3)
25	50	.5	.91	2.31	.37	1.13	.7	1.11	.4	.82	.58	.62	21 t2_R_B1C1 (mark 60.8)
12.5	25	.5	.86	1.85	.50	.91	-.5	.77	-.6	1.39	.62	.57	7 t1_R_C5 (mark 56)
25.5	51	.5	.84	1.69	.36	.93	-.3	.78	-.7	1.17	.66	.62	23 t2_R_C2D3 (mark 53.5)
25	50	.5	.81	1.48	.36	.92	-.4	.76	-.7	1.19	.65	.61	6 t1_R_C4D6 (mark 52)
25	50	.5	.71	.89	.47	1.03	.1	.84	.0	1.00	.78	.78	11 t1_R_E6F4 (mark 40.8)
12.5	25	.5	.68	.77	.47	1.19	.8	1.18	.7	.68	.39	.52	25 t2_R_D1 (mark 50.8)
12.5	25	.5	.66	.66	.51	.97	.0	.94	.0	1.04	.63	.62	10 t1_R_E5 (mark 43.3)
25.5	51	.5	.42	-.33	.33	1.11	.7	1.15	.8	.78	.44	.52	26 t2_R_D2E3 (mark 47)
12.5	25	.5	.35	-.60	.49	.83	-.7	.66	-.9	1.34	.68	.57	9 t1_R_D5 (mark 48.7)
25	50	.5	.30	-.85	.34	.89	-.6	.81	-.7	1.20	.62	.56	8 t1_R_D4E4 (mark 44.8)
12.5	25	.5	.20	-1.42	.49	1.10	.5	1.29	.8	.76	.49	.57	28 t2_R_E2 (mark 44.2)
12.5	25	.5	.15	-1.72	.49	.99	.0	.90	-.1	1.04	.59	.58	12 t1_R_F5 (mark 39.4)
25	50	.5	.15	-1.76	.36	1.08	.4	1.10	.4	.88	.57	.60	27 t2_R_E1F1 (mark 42.4)
12.5	25	.5	.13	-1.91	.48	.82	-.8	.70	-.7	1.36	.66	.56	30 t2_R_F3 (mark 40.2)
25	50	.5	.04	-3.17	.36	.83	-1.1	.58	-.9	1.35	.68	.61	14 t1_R_G4H4 (mark 34)
25	50	.5	.04	-3.21	.34	.97	-.1	.94	.0	1.05	.53	.52	29 t2_R_F2G3 (mark 38.8)
12.5	25	.5	.03	-3.37	.43	1.09	.7	1.17	.7	.56	.29	.38	31 t2_R_G1 (mark 36.4)
25	50	.5	.03	-3.57	.34	1.13	.9	1.07	.3	.73	.48	.54	13 t1_R_F6G6 (mark 37.3)
25	50	.5	.02	-4.01	.34	.89	-.7	.79	-.7	1.23	.62	.56	32 t2_R_G2H2 (mark 34.9)
12.5	25	.5	.01	-4.26	.54	1.13	.5	1.20	.5	.85	.62	.67	17 t1_R_H6 (mark 32)
12.5	25	.5	.00	-5.49	.62	1.05	.2	1.12	.4	.95	.74	.76	15 t1_R_G5 (mark 36)
12.5	25	.5	.00	-6.13	.53	1.16	.6	1.08	.3	.83	.60	.65	16 t1_R_H3 (mark 29.3)
12.5	25	.5	.00	-6.78	.57	.80	-.7	.45	-.7	1.32	.77	.70	33 t2_R_H1 (mark 30.1)
12.5	25	.5	.00	-6.78	.57	1.08	.3	.94	.1	.93	.68	.70	34 t2_R_H5 (mark 31.7)
12.5	25	.5	.08	(-2.42	1.84)	Minimum					.00	.00	24 t2_R_C6 (mark 57.5)
17.7	35.4	.5	.48	-.07	.51	1.01	.1	.91	-.1		.59		Mean (Count: 34)
6.2	12.4	.0	.41	3.97	.26	.12	.6	.26	.6		.16		S.D. (Population)
6.3	12.6	.0	.42	4.03	.26	.12	.6	.26	.6		.16		S.D. (Sample)

With extremes, Model, Populn: RMSE .57 Adj (True) S.D. 3.93 Separation 6.91 Reliability .98
With extremes, Model, Sample: RMSE .57 Adj (True) S.D. 3.99 Separation 7.01 Reliability .98
Without extremes, Model, Populn: RMSE .48 Adj (True) S.D. 3.98 Separation 8.29 Reliability .99
Without extremes, Model, Sample: RMSE .48 Adj (True) S.D. 4.04 Separation 8.42 Reliability .99
With 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.: 32 significance (probability): .44