CAMBRIDGE ASSESSMENT

Standard maintaining by expert judgment on multiple-choice tests: a new use for the rank-ordering method

Milja Curcin, Beth Black and Tom Bramley

Paper presented at the British Educational Research Association annual conference, University of Manchester, September 2009.

Research Division
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG

curcin.m@cambridgeassessment.org.uk
black.b@cambridgeassessment.org.uk
bramley.t@cambridgeassessment.org.uk

www.cambridgeassessment.org.uk

Abstract

The Angoff method for determining pass marks on multiple-choice tests is widely used in North America, Australia, and in the UK. It involves experts estimating the difficulty of multiple-choice questions for 'minimally competent' candidates (MCCs). However, as a standard setting method, it has no explicit mechanism for standard maintaining, i.e. keeping the pass mark at the same standard session on session. Therefore, there is a need to explore judgemental methods of standard maintaining for multiple-choice tests in situations where the requirements for statistical equating and linking are not met.

This paper investigates the use of the rank-ordering method (Bramley, 2005) as an explicit standard maintaining method based on direct comparison of the difficulty of multiple-choice questions from different examination sessions. The rank-ordering exercise was conducted twice on two OCR vocational qualifications (Certificate of Professional Competence in Road Haulage and Passenger Transport – CPC, and Award in Administration – AinA) that normally use the Angoff method for the same purpose so that the outcomes of the two methods could be compared. Each judge was given several packs of four questions (two from each session). Their task was to place the questions in each pack in rank order of perceived difficulty, thinking in terms of candidates in general for a given qualification, rather than just the MCCs. There was no access to performance data.

By fitting a Rasch model which estimates relative difficulty for each question based on the judges' rank orders, we obtained a common scale of 'perceived difficulty' on which to compare the two tests. This allows 'test equating' so that an equivalent pass mark can be set for the current session.

We found that the equating results based on rank-ordering judgements cross-validated the results of the Angoff procedure. However, a detailed analysis revealed that rank-orders of item difficulty based on the CPC rank-ordering judgements correlated poorly with the rank-order of empirical facilities while these correlations were much better for AinA. Nevertheless, both sets of judgements were shown to be reliable, exhibiting a high degree of consistency and agreement between judges. It was concluded that though sometimes deficient, the rank-ordering judgements represented judges' genuine view of question difficulty. While these results are clearly in need of replication, we conclude that since rank-ordering does not require 'precise' judgments of each question's difficulty, the correctness level at or above those observed for AinA in the current exercise would probably be acceptable as a valid basis for test equating. If such judgements could be consistently obtained for each session, the rank-ordering method would be a more defensible choice for standard maintaining owing to its conceptual and other advantages.

Introduction

In the context of criterion-referenced/related testing, once appropriate performance standards have been set for the first time for a new qualification to enable fair distinctions between, for example, competent and not-yet-competent candidates ('pass/fail' distinction), they then need to be maintained session on session. Various factors that could affect the pass marks and cause them to change have to be taken into account, not least test difficulty changes. Unless certain requirements for statistical equating are met,[1] standard maintaining often requires judgemental methods. However, few judgemental methods currently used for this purpose incorporate explicit mechanisms for directly comparing aspects of different examination sessions (e.g. test difficulty), though this is probably the most important aspect of any defensible standard maintaining method.

This study investigates the use of the rank-ordering method as an explicit standard maintaining method based on direct comparison of multiple-choice question (MCQ) difficulty from different examination sessions. This allows 'test equating' by expert judgement based on perceived test difficulty so that an equivalent pass mark can be set for the current session. The study was conducted on two OCR[2] vocational qualifications that normally use an iterative Angoff procedure for the same purpose (Angoff, 1971), which enabled the comparison of the outcomes of the two methods. Iteration attempts to minimise the impact of the inherent imprecision of human (even expert) judgement, in order for the ensuing decisions to be defensible.

According to OCR guidelines, after each live session, five to eight awarders (usually teachers or question setters) initially individually estimate the percentage of a group of minimally competent candidates (MCCs - those candidates with sufficient skills to only just achieve a pass) who would select the correct answer for each question on the test (pre-Angoff procedure). This estimate is a proxy for the question's level of difficulty. The estimates can be amended following a discussion informed by performance data (question facilities,[3] candidate score distribution). The average of all final estimates is then calculated, giving the recommended pass mark for the test. In light of this recommendation as well as the historical and impact data, the endorsement panel determines the final pass mark.

In place of an explicit comparison mechanism focusing on question difficulty, the Angoff method assumes that the judges can consistently conceptualise the MCCs within and between sessions. Indeed, the assumption of consistency in a conceptualisation of MCCs is the very crux of how Angoff can function as a standard maintaining method, despite research evidence suggesting that an MCC's performance is actually difficult to conceptualise (e.g. Impara and Plake, 1998; Bouriscot and Roberts, 2006). It is also questionable whether judges are able to estimate absolute question difficulty accurately. While judges are generally able to rank order questions in terms of difficulty, as moderate correlations between difficulty estimates and empirical facilities indicate, around 50% of their difficulty estimates for individual questions tend to be inaccurate (Brandon, 2004; Idle, 2008). Another very important, and generally contentious, issue is that of 'contamination' of professional judgement by discussion and exposure to performance data, which could undermine its independence, reducing the method's defensibility and face validity (e.g. Newton, 2000).

The rank-ordering method trialled in this study attempts to address some of the abovementioned issues. The method (Bramley, 2005, cf. Thurstone, 1931) is an extension of the paired comparisons method for capturing relative judgements of non-physical attributes, e.g. 'seriousness of crime' that cannot be otherwise measured (Thurstone, 1927). Repeated comparisons of entities (e.g. scripts) containing different degrees of a property/trait (e.g. quality) yield a single scale for that trait and the location of each entity on that scale in terms of how

[1] This is possible only if there are common items or common candidates between sessions (see e.g. Kolen and Brennan, 2004).

[2] OCR (Oxford, Cambridge and RSA) is one of the major awarding bodies in the UK

[3] In classical test theory, facility is calculated as the ratio between the mean mark and the maximum mark for a question, indicating question difficulty.

much of the trait it is judged to possess. Rank-ordering produces a similar outcome through a more efficient procedure since rather than carrying out repeated paired comparisons, judges rank several sets (packs) of, for instance, 10 scripts, which gives 45 paired comparisons per pack on which the scale of quality is based.

Several rank-ordering exercises have been conducted to date in order to investigate the method's validity and reliability in comparison with other standard maintaining activities, in different contexts, and using different designs. Most previous studies have compared the rank-ordering outcomes with those of awarding (e.g. Black and Bramley, 2008a; Gill, Bramley and Black, 2007; Gill and Bramley, 2008).[4] In these studies, the judges made holistic judgements of script quality without access to original marks for tests containing mainly open-ended questions. The derived measure of script quality correlated very well (0.8 or 0.9) with the original marks while the method has proven robust, rigorous and capable of being cross-validated.

However, script quality judgements are less appropriate for standard maintaining on objective tests, as there is less differentiating content to help judges form ideas about the perceived quality of the script. Furthermore, an important issue with standard maintaining based on script quality judgements is that these are, arguably, a proxy for getting at change in test difficulty, which is the core pre-requisite of standard maintaining (Bramley and Black, 2008). Using the rank-ordering method to elicit judgments about question difficulty should meet this pre-requisite more directly.

As an explicit standard maintaining method, in theory rank-ordering also has several advantages over standard setting methods such as the Angoff:

(i) Rank-ordering allows direct comparison of tests from two or more sessions.
(ii) Being based on comparison of one question with another, rank-ordering does not require the judges to conceptualise a specific competence level of candidates, nor for all the judges to conceptualise the same competence level.
(iii) Rank-ordering questions without access to performance data is a 'pure' way of capturing expert judgement. It helps to keep different sources of standard maintaining evidence separate.
(iv) The method offers the possibility of testing the judgements not only for accuracy but also for consistency, which contributes to its face validity.

However, despite all these potential advantages, we are yet to establish whether judgements of question difficulty would indeed improve compared with Angoff if obtained through the rank-ordering method. One reason to believe they would is that rank-ordering requires relative judgements, which, based on the evidence from the relevant literature (e.g. Laming, 2004; Gill and Bramley, 2008), should be easier to make than absolute judgements such as Angoff estimates. On the other hand, judging question difficulty for candidates seems to be a more complex task for judges than, for instance, judging script quality. It is not easy to conceptualise in a straightforward way what constitutes question difficulty, with a multitude of factors playing a part (see e.g. Fisher-Hoch, Hughes and Bramley, 1997; Pollitt, Ahmed and Crisp, 2007). Also, while teachers or examiners often have to make script quality judgements, for instance in marking, making explicit question difficulty judgements is not often required. For these reasons, we could perhaps expect somewhat less accurate and reliable[5] results from rank-ordering questions in terms of difficulty than rank-ordering scripts in terms of quality.

[4] The method also has the potential for use in vertical equating (Black and Bramley, 2008b), GCSE Design and Technology portfolio assessment (Kimbell et al., 2007) and inter-board comparability studies.

[5] Throughout this paper we differentiate between the following two characterisations of judgements (both difficulty estimates and difficulty rank-orders): (a) *accurate*: refers to judgements that agree with empirical facility values, either absolutely or in terms of rank-order of difficulty; (b) *reliable*: describes agreement between judges regarding their difficulty estimates or rank-orders (includes consistency in judgements between sessions in which the same tests were compared)

Method

Qualifications and judges

The qualifications chosen for this rank-ordering exercise were:

- OCR Level 3 Certificate of Professional Competence (CPC) in Road Haulage and Passenger Transport, Unit 1 – Understanding the Legal and Business Context for Road Transport Operations (05598)
- OCR Level 2 Award in Administration (AinA), Unit 1 – Identifying Administrative Functions (03790)

Each test contains 30 MCQs. The results are graded as Pass or Fail, with notional pass rates at approximately 60% of the available marks, but may vary somewhat as a result of the Angoff recommendations and endorsement panel decisions.

The exercise was conducted in four stages. This design allowed direct comparison of questions from a previous session with those from the current live session of each test and also enabled a comparison of judgement consistency on the December (CPC) and the April tests (AinA), which were used twice in stages 1/2 and 3/4 respectively (Table 1).

Table 1: Stages of the rank-ordering exercise

Qualification	Stage	Session		
CPC	1	September 08	December 08 ***(live)***	
	2		December 08	March 09 ***(live)***
AinA	3	November 08	April 09 ***(live)***	
	4		April 09	June 09 ***(live)***

The same judges that normally take part in the Angoff panel were recruited. Their number varied between the stages and qualifications (see below), but most of them were the same between stages.

Pack design

Packs of four

While previous rank-ordering exercises used packs of ten scripts (e.g. Bramley, 2005), six (Black and Gill, 2008) or three (Raikes et al., 2008), the current study used packs of four questions. In the current context, this appeared to be cognitively easier than, for instance, packs of ten or six, while giving an even number of questions per session per pack. As Figure 1 shows, each pack of four questions produces 6 paired comparisons.

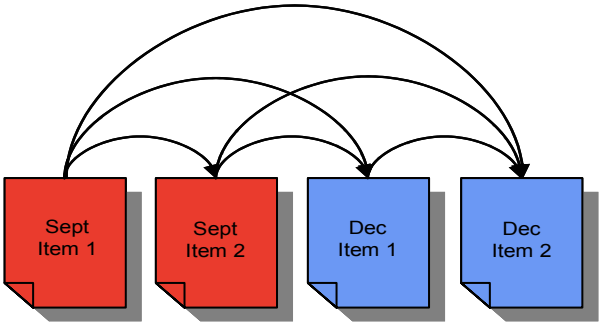


Figure 1: Comparisons generated from rank ordering four questions

Overall pack design

Pack design varied slightly over the stages as a function of the number of judges that we were able to recruit. In each stage, the overall aim was to obtain the desirable ≈50% of the possible number of comparisons (n=1770) obtainable from 60 questions on two tests. In addition, in order

to obtain parameter estimates sufficiently accurate for linking the two scales of difficulty together, it was necessary for each question to enter into several comparisons across judges and packs in each stage (at least 16 times, see Kimbell et al., 2007). Given these general constraints, the distribution of questions, judges and packs for each stage was as follows:

Table 2: Distribution of judges and packs in all stages

Qualification	Stage	Judges	Packs
CPC	1	6	25
	2	7	25
AinA	3	4	30
	4	5	30

The guiding principles for the pack design of the study were:

- to ensure each judge's exposure to each question at least once
- to minimise the number of times a judge saw any one question – no judge saw any question more than twice
- for each question to be seen an equal number of times across judges
- for each judge to have a unique set of packs and combination of questions
- to ensure no judge made the same paired comparison twice

Table 3: Summary of pack design for all stages

No. of paired comparisons generated per judge	150-180
Total no. of paired comparisons	720-1050
Frequency of exposure to same question per judge	Min 1, Max 2
Frequency of same question exposure across judges	8-10
No. of comparisons each question enters across packs and judges	24-30

The major part of the pack design was generated by random allocation and then tweaked to meet all the abovementioned constraints (see Appendix A for a portion of the matrix displaying the way the questions were allocated to packs and judges in stage 2). The questions from each test were scanned, presented on individual sheets of paper together with the correct answer, allocated a unique ID (e.g. A01, B29) according to the session they belonged to (e.g. A for September, B for December) and combined into packs according to the design described.

Procedure

As the exercise was conducted during live sessions, for security reasons the packs were sent to the judges on the day the exam was sat to reach them the following day. The judges were given up to 2 weeks to complete the task at home and return the materials before they received questions from the live session for pre-Angoff. This meant that all the materials were returned before the judges became familiar with actual candidate performance on the live test. The task materials for each judge consisted of:

- 25/30 packs of questions
- A recording form for each pack
- A summary recording form for all packs
- Instructions (see Appendix B)
- FAQs (see Appendix C)

The judges were informed that each pack contained two questions from the previous and two questions from the current live session. They were asked to place the questions in rank order, from most difficult to easiest, based on their professional judgement of the questions' relative difficulty. It was clearly stated that the current exercise was not a version of pre-Angoff and that they were to *compare* the questions in each pack with one another, and not estimate individual question difficulty. The judges were asked to think of *candidates in general* for this qualification, or focus on a familiar group of candidates, rather than MCCs. Tied ranks were not allowed.

Data analysis

The ranks obtained were converted into paired comparisons and analysed by fitting a Rasch paired-comparisons model (Andrich, 1978) using the FACETS software (Linacre, 2005):

$$ln[P_{ij} / (1-P_{ij})] = \theta_i - \theta_j$$

Where P_{ij} = the probability that question *i* beats question *j* in a paired comparison
and θ_i = the measure for question *i*
and θ_j = the measure for question *j*

The analysis produces a latent trait scale (a common scale of difficulty), and its unit 'logit' or 'log-odds unit' denotes the amount of difficulty (measure) each question was perceived to have.

The next phase of the analysis was the test equating. Once the perceived difficulties of the questions in the two tests have been calibrated onto the same scale by the rank-ordering method, they can be treated in the same way as a calibrated item bank created by the more usual methods of pre-testing, anchoring and equating. The Test Characteristic Curve (TCC) is a plot of expected score on test against ability. The expected score on the test is the sum of the expected scores on each question for a candidate of a given ability. For a dichotomous question, the expected score is the probability of success (P) of person *n* on question *i*, as given by the equation for the Rasch model ($ln[P_{ni} / (1-P_{ni})] = \theta_n - b_i$), where θ_n represents examinee ability and b_i represents question difficulty. The expected score on the test (known as the 'true score') is the sum of these probabilities across the questions on the test:

$$TS_j = \sum_{i=1}^{N} P_i(\theta_j)$$

where: TS_j is the true score for examinees with ability level θ_j.
i denotes a question and $P_i(\theta_j)$ is obtained via the abovementioned Rasch model.

If the question difficulties are known, then the expected test score for a given level of ability (or the ability corresponding to a given expected test score) can be derived by iteration. In this study, we used the 'goal seek' function in MS Excel. The abilities corresponding to each possible raw score on the test from 1 to 29 were obtained by this method.[6] TCCs can then be plotted based on these results. If the TCCs for the two tests are plotted on the same graph, it is possible to find the pass mark on the live test corresponding to a given pass mark on the previous test.

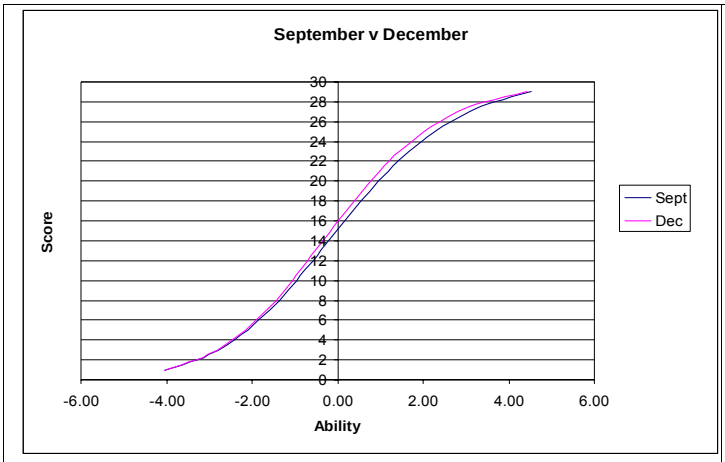
An evaluation of the rank-ordering judgements underlying the equating results was conducted by investigating the fit of the data to the Rasch model (item and judge residuals[7]; separation and separation reliability[8]) and the correlation between the measures of difficulty obtained in the rank-ordering exercise and empirical facilities for the relevant session.[9] Difficulty measure is expressed on a logit scale, ranging from negative (easier) to positive values (more difficult), while facility values range from 0 to 1 and can be expressed in percentages, with lower values for more difficult questions, and higher values for easier questions.

[6] The scores necessarily range from 1 to 29, rather than from 0 to 30, as it is impossible to estimate measures for extreme scores.
[7] Residuals lower than 2.5 were considered acceptable.
[8] Separation is a measure of the spread of the estimates compared to their precision and is calculated using the adjusted standard deviation over the root mean square standard error. A higher value of separation suggests we are reasonably sure that the measures are 'separated' from one another, either due to a large overall standard deviation or because the estimates are more precise (have a lower standard error). Separation reliability is the ratio of true variance to observed variance, which indicates the proportion of the variation in the measures which can be attributed to differences between the items. It is analogous to Cronbach's Alpha in traditional test theory. See Wright and Masters (1982).
[9] Rank-ordering is a 'strong' method in that it can be invalidated if either (a) measures and facilities do not correlate, and (b) the data fail to fit the model and/or fail to create a meaningful scale (Bramley and Black, 2008). These problems can occur independently and, if either is detected, this calls into question the validity of the final equating.

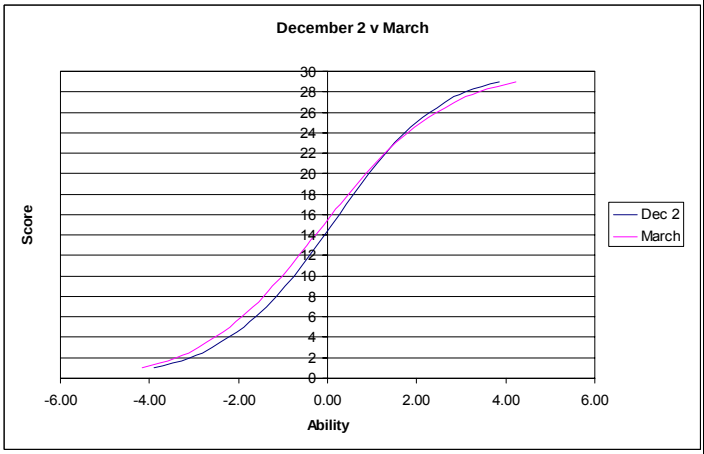
Results

Determining the pass mark

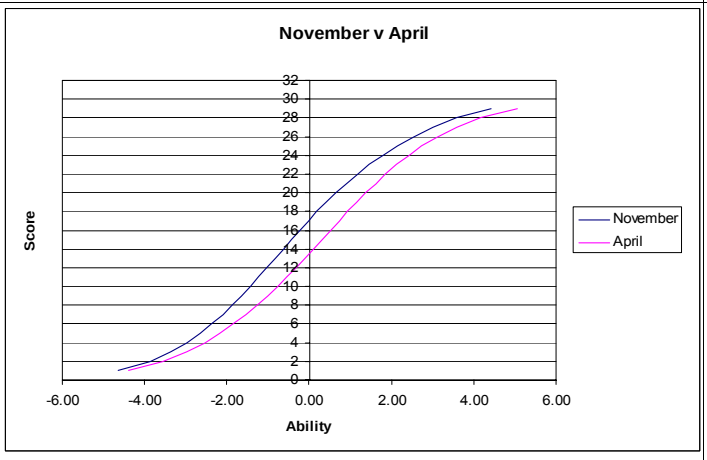
Using the equating procedure described above, we obtained the graphs in Figure 3 below, showing the TCCs for the two tests in each stage together. This allowed us to determine the pass mark on the live test in relation to the pass mark on the previous test.



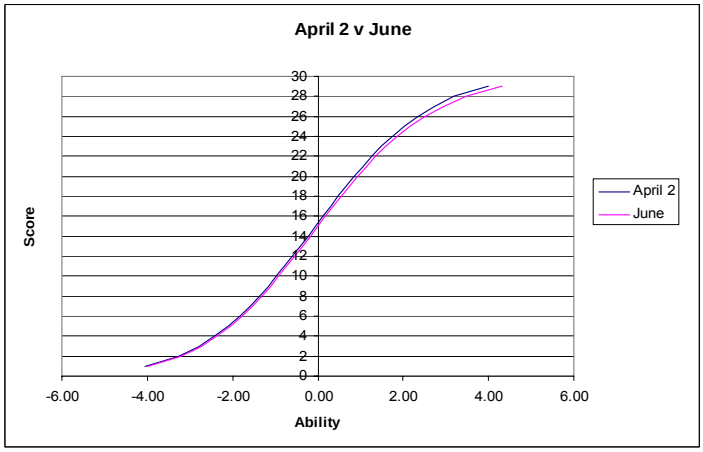
Score	Sept ability	Dec ability
20	0.95	0.78
21	1.18	0.99



Score	Dec 2 ability	March ability
21	1.11	1.07



Score	Nov ability	April ability
19	0.41	1.15
20	0.65	1.37
21	0.90	1.61
22	1.17	1.86



Score	April 2 ability	June ability
19	0.66	0.73
20	0.85	0.93
21	1.06	1.14
22	1.27	1.36

Figure 3: Test equate for all stages

According to the rank-ordering judgements, the December test for CPC stage 1 was perceived as easier than the September test. In other words, a candidate of equivalent ability would have achieved the score of 21 in December, and 20 in September. In stage 2, the CPC judges saw the two tests as equivalent in terms of difficulty, at least in the average score region (around 20-24 marks). For AinA, there was more variability in test difficulty between sessions according to the rank-ordering judgements. In stage 3, the April test was judged as more difficult than the November test, while in stage 4 the tests were perceived as equally difficult.

Table 4 summarises the pass marks based on the initial and final Angoff recommendations, the endorsed pass marks[10] and the rank-ordering pass mark.

Table 4: Pass mark recommendations based on different judgemental procedures

Qual.	Stage	Session	Pass mark		
			Endorsed (pass rate %)	Angoff final (pass rate %)	Rank-ordering (pass rate %)
CPC	1	Sept 08	20 (57.65)	21 (50.23) ×	
		Dec 08	21 (65.28)	21 (65.28) ✓	21 (65.28) ✓
	2	Dec 08			
		Mar 09	21 (62.61)	21 (62.61) ✓	21 (62.61) ✓
AinA	3	Nov 08	22 (66.67)	21 (80.26) ×	
		April 09	20 (61.42)	21 (48.82) ×	19 (73.23) ×✓
	4	April 09			
		June 09	21 (62.33)	20 (73.02) ×	19 (80.47) ×

In stages 1 and 2, the pass mark based on rank-ordering matched the endorsed pass mark perfectly, while the final Angoff recommendation for September diverged from the endorsed pass mark by one mark. The situation is more complicated for stages 3 and 4. In stage 3, the endorsement panel lowered the pass mark for April to 20 (from the November level of 22), consistent with a belief that the April test was more difficult. This was also reflected in impact information. If the pass mark for April remained at 22, the pass rate would have been just 32.28%, as opposed to 61.42% with the pass mark of 20 (see Table 4). The rank-ordering judgements went in the same direction, but one mark lower. In contrast, the Angoff panel did not suggest any change in pass mark between November and April. In stage 4, according to the rank-ordering judgements, the pass mark for June remained at 19. However, the endorsed pass mark for June was raised by one mark to 21, suggesting that the test in this session was easier than the one in April. Conversely, the final Angoff panel recommendation of 20 suggested that they perceived the June test as more difficult than the April test.

It is difficult to adjudicate which of the three recommendations in stages 3 and 4 is the 'correct' one. Clearly, in stage 3, the rank-ordering recommendation was closer to the endorsed pass mark than the final Angoff recommendation. In stage 4, both the rank-ordering and the Angoff recommendations diverged from the endorsed pass mark, and neither method suggested raising the pass mark, contrary to the endorsed decision. While all the pass marks are very close, even single mark changes are usually accompanied by large changes in pass rates in AinA. In such qualifications, impact information is likely to play an important role in pass mark decisions irrespective of the judgemental method used.

Overall, these results show the rank-ordering method in a favourable light, in that in most cases it cross-validated the endorsed and Angoff pass marks. Furthermore, this was achieved without discussion, access to performance data and impact considerations though the rank-ordering method does not exclude the possibility of using performance and impact data in later stages if this is needed to inform the final decision. However, in this method, the judgemental outcome is completely independent, unlike in the case of the full Angoff procedure, where it is difficult to dissociate expert judgement from other influences.

Evaluation of rank-ordering judgements

In order to evaluate the rank-ordering judgements on which the abovementioned equating procedure is based, we consider the overall fit of the data to the Rasch model, measure-facility correlations, judgement consistency, and how the rank-ordering judgements compare with the Angoff judgements. Although the purpose of the rank-ordering exercise is to provide a mechanism for comparing two or more sessions within each pack, the analyses are necessarily

[10] Note that the determination of the final pass mark at this stage relies to a great extent on impact information, i.e. the percentage of the candidates that would pass depending on which pass mark is chosen.

conducted *within* session since we cannot assume that the cohorts between sessions (and therefore empirical facility values) are directly comparable. However, it is plausible to assume that these results are generalisable to those between sessions.

Fit

The rank-ordering data fit the model well in all stages, with separation ranging from 2.72 to 2.82, separation reliability of .88 or .89,[11] and item residuals generally within acceptable limits. In the current context, high separation reliability indicates that the judges perceived a similar scale of difficulty for the items in question, i.e. that their judgements were reliable. There was a small number of misfitting judgements (n=9-12) in each stage although overall judge fit to the model was good, additionally indicating a high level of agreement between judges. These values are slightly lower but overall comparable with those obtained in previous rank-ordering exercises.

Measure-facility agreement

A measure-facility correlation for corresponding questions gives us an indication of the extent to which the respective rank-orders of these values agree. A strong *negative* correlation would show a high level of agreement, indicating that the judges were good at judging relative question difficulty, giving us confidence that the equating results are based on sound expert judgement.

However, the analysis of the measure-facility agreement for the two CPC stages showed that these correlations were very weak (see Table 5 and Figures 4a and 4b below), and even in the wrong direction in one case. In other words, the questions that turned out to be easy for candidates (higher facilities) were judged as difficult (gained a higher measure of perceived difficulty) in the rank-ordering exercise. A slight improvement is apparent in stage 2 with the correlation for December in the right direction, though somewhat lower correlation for March.

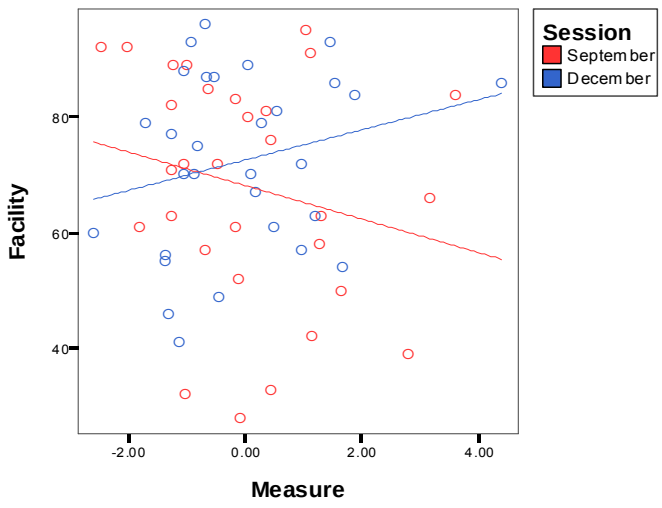
In stages 3 and 4 (AinA) we obtained better results (Figure 4c and 4d), especially for June, where the measure-facility correlation was higher than in any of the preceding stages as well as highly statistically significant (i.e. significantly different from zero). The correlations for the other AinA sessions were also in the right direction and appreciably higher than for CPC, suggesting that AinA judges were better at judging question difficulty. There was again variability between sessions in each AinA stage, and an improvement in stage 4 compared to stage 3.

Table 5: Measure-facility correlations for all sessions

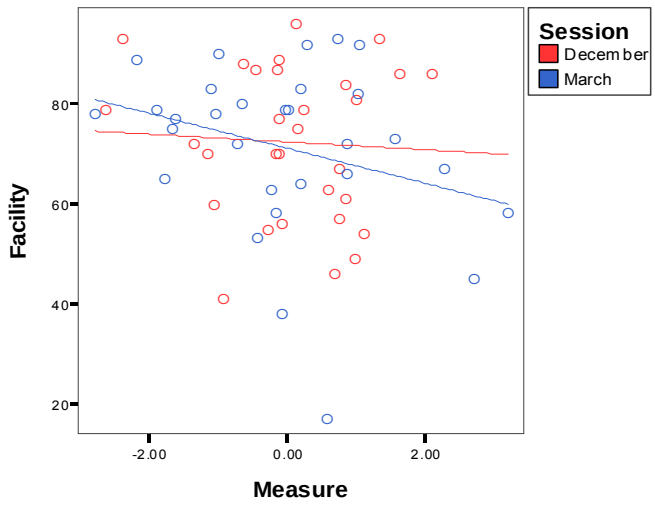
Qualification	Stage	Session	Spearman correlation
CPC	1	September	-.260
		December	.230
	2	December	-.064
		March	-.178
AinA	3	November	-.459*[12]
		April	-.337
	4	April	-.343
		June	-.629**

[11] These separation reliability coefficients are likely to be overestimates because of violation of local independence in the rank-ordering method (Linacre, 2006).

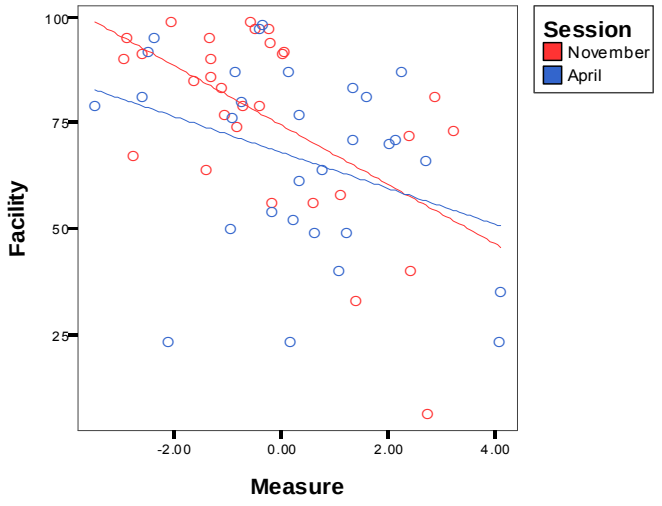
[12] In this and other tables presenting correlations, * represents significance at 0.05 level, and ** represents significance at 0.01 level (two-tailed).



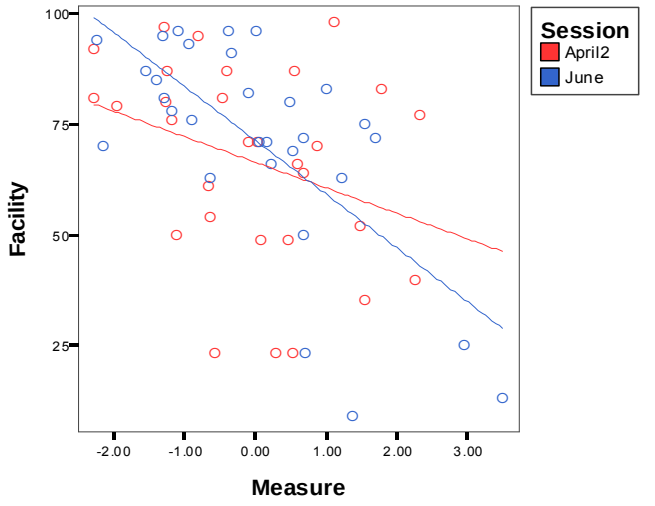
a. Stage 1 - CPC



b. Stage 2 - CPC



c. Stage 3 - AinA



d. Stage 4 - AinA

Figure 4: Graphs summarising measure-facility correlations

It is disconcerting that the rank-ordering judgements failed to have a good agreement with the empirical facilities, especially in the case of CPC (stages 1 and 2). While the AinA results are perhaps acceptable, the CPC results actually undermine the equating results discussed previously. Such low correlations are unexpected considering the results of the previous rank-ordering studies conducted on scripts rather than question difficulties, where mark and measure correlations were very good. This discrepancy is certainly larger than expected, especially for CPC, even considering our initial expectation that the current exercise might yield somewhat worse results as question difficulty is probably more difficult to judge than script quality. In order to further evaluate the method and uncover whether this result was perhaps an artefact of the current methodology, or a consequence of some judges' genuine inability to judge question difficulty, we present an investigation of the following aspects in the next three sections:

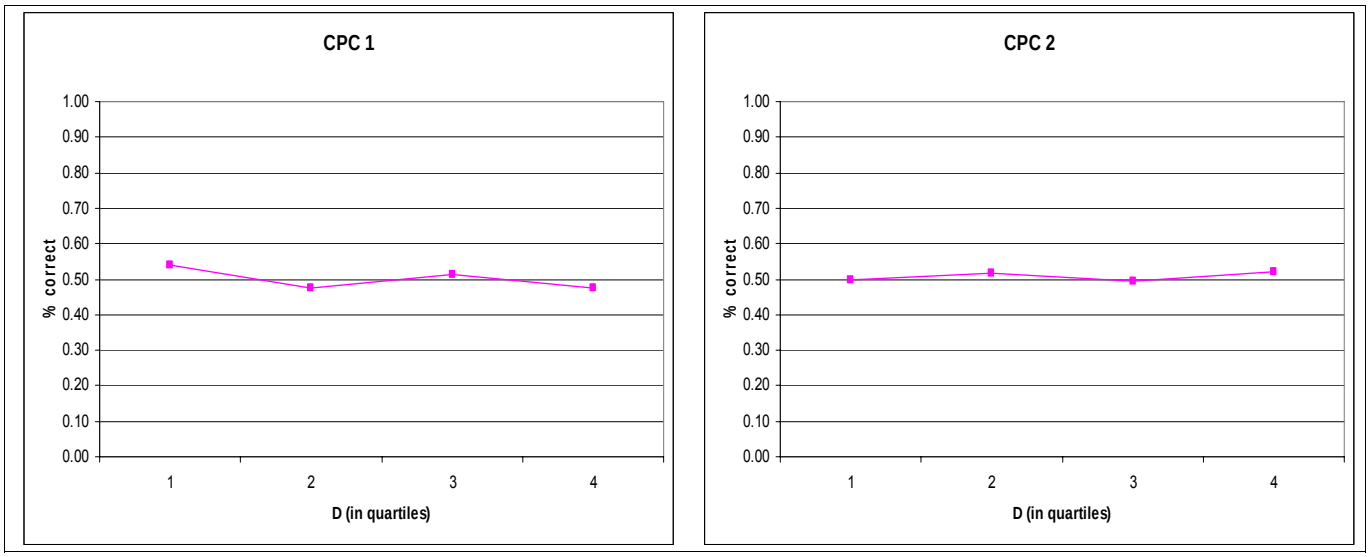
- The potential influence of the pack design
- Consistency of the rank-ordering judgements between overlapping sessions
- Whether rank-ordering managed to capture any genuine aspects of expert judgement and how the effectiveness of relative and absolute judgement compared as revealed in a comparison of the rank-ordering outcomes with the outcomes of the Angoff procedure

Have aspects of pack design affected the judgemental outcome?

A possible explanation for the particularly poor relationship between measures and facilities for CPC would be that the pared down pack design (only 4 questions per pack and a relatively small number of judges) might not have yielded enough comparisons to absorb the impact of misfitting judgements. This is, however, unlikely considering good fit of the data from both CPC stages to the Rasch model, as well as the small number of misfitting judgements. In addition, a similar pack design was used for AinA, resulting in much better measure-facility correlations.

Another possibility is that due to random allocation of questions to packs, questions may have been (too) close in terms of facility in some packs and further apart in others. In the former case, it would presumably have been more difficult for the judges to tell question difficulty apart and rank-order questions correctly. We explored the judges' sensitivity to facility differences between questions in their packs by plotting facility difference ranges (D, averaged per quartile) within each judge's pack and session against the percentage of correct judgements for the corresponding range (c.f. Gill and Bramley, 2007 for a similar analysis). This is shown in Figure 5 and Table 6 below. We would expect to see increasing correctness rate with increasing facility difference if this aspect had an effect on the rank-ordering judgements.

The charts below show that facility differences had no discernible impact on the CPC judges' judgement correctness[13], while having a differential effect on judgement correctness of AinA judges. In the current case, this suggests that the AinA judges were generally more aware of question difficulty levels than the CPC judges, finding them more obvious in packs where facility differences between questions were greater. This situation indicates that the difference in the CPC and AinA rank-ordering results was not related to this aspect of pack design either but to a genuine difference in the quality of the rank-ordering judgements between the two sets of judges.



[13] Correctness rate refers to the proportion of judges' ranks matching the ranks based on the empirical facilities within each session and pack.

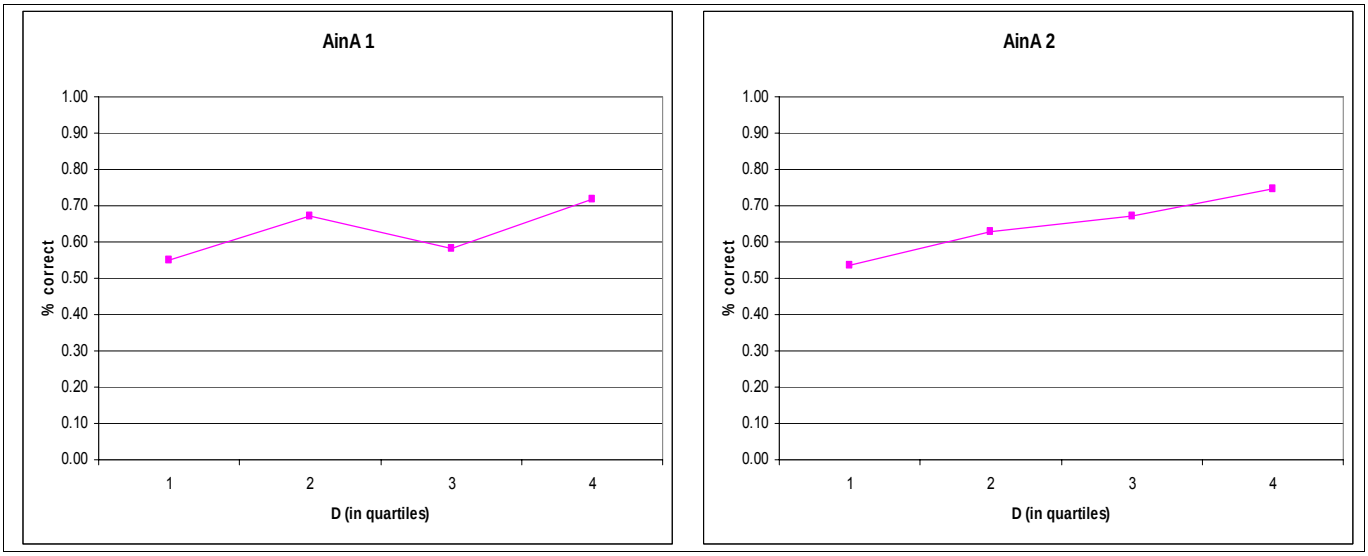


Figure 5: Percentage of times rank-order within packs agrees with the original facilities order against rising facility differences

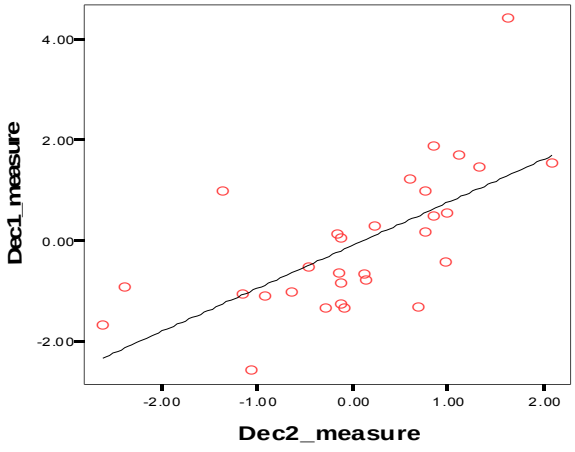
Table 6: Summary of the facility ranges within each quartile and corresponding correctness rates

Qualification	Quartile	Facility range	% correct	No. of judgements
CPC stage 1	1	0-8	0.54	72
	2	9-17	0.47	76
	3	18-29	0.51	74
	4	30-64	0.47	78
CPC stage 2	1	0-6	0.50	80
	2	7-14	0.52	89
	3	15-26	0.49	91
	4	27-75	0.52	88
AinA stage 3	1	0-9	0.55	60
	2	10-19	0.67	58
	3	20-35	0.58	62
	4	36-85	0.72	60
AinA stage 4	1	0-9	0.54	71
	2	10-20	0.63	81
	3	21-39	0.67	73
	4	40-87	0.75	75

Judgement consistency

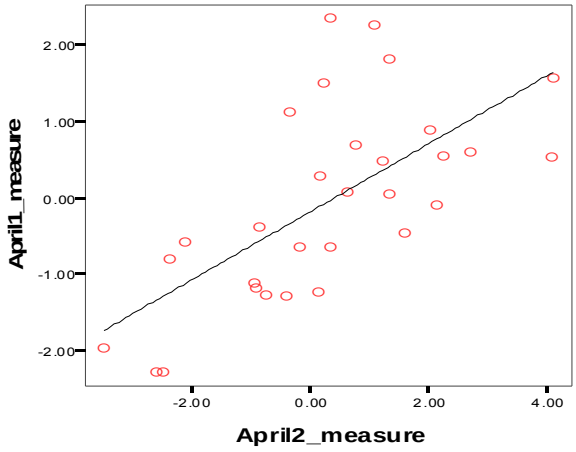
A comparison of the sets of measures obtained for the common paper between the two stages for each qualification gave an indication of judgement consistency. A poor relationship would suggest that the judges did not have a clear, consistent view of each question's relative difficulty, and would indicate untrustworthy judgements. This provides another way of evaluating the reliability of the rank-ordering method (in addition to separation and separation reliability).

Significant correlation between the two sets of measures for the two CPC December and the two AinA April sessions (Figures 6a and 6b) suggest that the judges were quite consistent in their judgements of the difficulty of the same questions. This consistency indicates that the judgements elicited via rank-ordering are reliable and reflect judges' actual perception of the relative question difficulty. This finding gives additional support to our conjecture that the low measure-facility correlations obtained for CPC are not a consequence of the current pack design, or a relatively small number of judges, which would likely have caused the absence of strong relationship between the two sets of measures.



Pearson correlation	
Dec 1 vs. Dec 2	.656**

a. December 1 vs. December 2 (CPC)



Pearson correlation	
April 1 vs. April 2	.658**

b. April 1 vs. April 2 (AinA)

Figure 6: Relationship between the measures of difficulty obtained in the two December and two April sessions

Comparison with the Angoff judgements

An indirect comparison of the rank-ordering and Angoff judgements from the corresponding sessions should indicate which outcome is a closer match for empirical facilities, and whether the two methods capture similar aspects of professional judgement. We would predict, based on previous research, that the method relying on relative judgements should give better results. For all comparisons in this section, we use *initial* Angoff estimates since these are made independently of other panel members and are not informed by performance data.

The correlation between initial Angoff estimates and the relevant empirical facilities for each session gives an indication of how well the rank-order of question difficulty derived from these absolute estimates agreed with the facilities,[14] providing a baseline for comparison with the rank-ordering method outcomes. Table 8 shows that these estimates (averaged across judges) are significantly correlated with the corresponding empirical facilities for all sessions except for CPC December and March where there is a moderate (non-signficant) correlation. Variability within and between qualifications is also present, the AinA judges appearing to be better overall at estimating relative question difficulties.

A comparison of the Angoff and rank-ordering correlations shows big differences between the two qualifications. For CPC, the correlations suggest that the Angoff judgements were closer to the empirical facilities than the rank-ordering ones. These differences were much less pronounced for AinA sessions. While these results are not clear-cut, they suggest that, despite being based on relative rather than absolute judgements, the rank-ordering method failed to elicit judgements that are generally better correlated with the facilities than those elicited by the Angoff method.

[14] Note, though, that the correlation between empirical facilities and Angoff estimates only gives us an indication of how well the judges were able to perceive a rank-order of item difficulty. It does not reveal how precise their judgements regarding individual items were (cf. Idle, 2008).

Table 8: Correlations between facilities and initial Angoff estimates vs. rank-ordering measures

Qualification	Angoff session	Spearman correlation	Rank-ordering session	Spearman correlation
CPC	September	.613**	September	-.260
	December	.358	December – stage 1	.230
			December – stage 2	-.064
	March	.301	March	-.178
AinaA	November	.615**	November	-.459*
	April	.534**	April – stage 3	-.337
			April – stage 4	-.343
	June	.612**	June	-.629**

Interestingly, in each method the judgements for the CPC December and March sessions and for AinA April sessions were less in accord with empirical facilities than the judgements for the remaining sessions within each qualification. This suggests that rank-ordering captured relatively similar aspects of professional judgement as the Angoff method. Further evidence for this is provided by largely good correlations of the rank-ordering measures and initial Angoff estimates, showing that many of the rank-ordering and Angoff judgements were similar within each session (Table 9). This also supports our argument that the rank-ordering judgements obtained in the current study are not an artefact of pack design, but a genuine representation of judges' views regarding question difficulty.

Table 9: Correlations between rank-ordering measures and Angoff initial estimates

Qualification	Stage	Session	Spearman correlation
CPC	1	September	-.697**
		December	-.374*
	2	December	-.209
		March	-.844**
AinA	3	November	-.484**
		April	-.498**
	4	April	-.593**
		June	-.764**

Discussion

Perhaps the most striking outcome of the current rank-ordering exercise is the disparity between encouraging equating outcomes and generally questionable quality of underlying judgements. Though the pass marks based on rank-ordering matched the endorsement results quite well (perhaps even more consistently so than the final Angoff pass marks), the rank-ordering measure-facility correlations were mostly worse than pre-Angoff estimate-facility correlations (though the latter were themselves not as good as might be expected of a method that requires quite precise estimates of each question's difficulty). Such low measure-facility correlations were unexpected based on the outcomes of the previous rank-ordering exercises.

Whilst it is impossible to pinpoint absolutely why this is so, we have argued that the rank-ordering judgemental outcome has not been affected by current pack design. We also found that the judgements were consistent between overlapping sessions, suggesting that they represent the judges' actual perception of question difficulty. We could therefore probably attribute the abovementioned results to the judgements of relative difficulty being deficient in some cases.

We also observed differences between the two qualifications in terms of judgment quality, the AinA November and June session judgements reaching perhaps even acceptable level of agreement with the facilities. There was also a significant degree of variability between judges, which is also characteristic of the Angoff method. In addition, the two methods appear to tap into similar aspects of expert judgements, though perhaps to different degrees. However, there was

no compelling evidence of the relative judgements elicited through rank-ordering being more accurate than the Angoff absolute judgements except for the AinA June session.

At this point, we can only speculate why rank ordering failed to elicit consistently valid judgements about question difficulty when, in previous rank-ordering studies, there is good evidence of valid judgements about script quality. It is also difficult to explain why the rank-ordering judgements ended up being less accurate than the pre-Angoff ones in some cases.

Part of the reason may be a lack of familiarity with the rank-ordering procedure. A small practice effect observed between the two stages within each qualification gives some support to this possibility. This would suggest that some training prior to the rank-ordering exercise might be beneficial. Another possible influence on judge performance may be related to judge expertise. As previously discussed, judging question difficulty is perhaps a more complex (or less familiar) task than judging script quality. It would, therefore, probably be useful for any training sessions on rank-ordering to include some discussion on what constitutes question difficulty, in order to standardise judges in this respect as much as possible. Of course, the judges would probably benefit from this sort of training even in the context of the Angoff procedure. However, speculating further on these issues would not take us very far without further research.

The most encouraging result of this rank-ordering exercise is the equating outcome. The rank-ordering and equating procedures appear to have captured or distilled overall relative difficulty profiles of the tests in each stage, based on a certain number of correct difficulty differentiating judgements in order to relate these to abilities and derive the outcome on which to base the pass mark. At this stage, it is unclear how many such judgements is enough for an acceptable outcome, though the amount obtained from the CPC judges probably renders the equating outcome indefensible. However, as rank-ordering does not require ‘precise’ judgments of each question’s difficulty (thus recognising the inherently imprecise nature of expert judgement in this domain), the correctness level at or above those observed for AinA would probably be acceptable. If such judgements could be consistently obtained for each session, the rank-ordering method would be a more defensible choice for standard maintaining owing to its conceptual and other advantages discussed in the introduction.

An issue with the current approach, as well as the Angoff method, is the possibility of artefacts related to the ‘equating’ procedures used. In the Angoff procedure, there is a possibility that since the judges have a vague notion of the range of facilities for each test session, the average estimate converges on the average mark for a given test. This problem has been noted in the literature as regression towards the mean, or centre of the probability scale, (see e.g. Lorge and Kruglov, 1952; Schulz, 2006; Reckase, 2006). In rank-ordering, similar effects might arise in that even random rank-ordering judgements might result in similar pass marks once the tests are equated in the previously described way. Initial examination of this problem in connection with rank-ordering suggests that different simulated sequences of random rank-ordering judgements produce different equating results from those based on actual judgements, though further investigations are needed. Note, however, that the rank-ordering method provides a possibility of evaluating the judgemental outcome in several independent ways. The results of these evaluations could then either support or disprove the equating results. It is less clear how such an evaluation could be conducted in the context of the Angoff procedure.

While partly encouraging, the results of the current rank-ordering exercise are inconclusive. They require replication in the context of other qualifications that use MCQs or other kinds of objective questions, as well as using different pack designs. We also need to further investigate potential influences on rank-ordering judgements by controlling for the effects of training and judge expertise, as well as investigate ways of standardising the notion of question difficulty for the judges. Finally, although further investigation of the current application of the rank-ordering method would be very useful, it is also important to continue the search for alternative standard maintaining methods that incorporate direct comparison between consecutive sessions, which is perhaps the most important aspect of any valid standard maintaining method.

References

Angoff, W. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.) *Educational Measurement (2nd ed.).* American Council on Education, Washington, DC, 508-597.

Black, B. and Bramley, T. (2008a). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education, 23(3),* 357-373.

Black, B. and Bramley, T. (2008b). Using expert judgment to link mark scales on different tiers of a GCSE English examination: a rank-ordering method. Presentation given at the 3rd annual meeting of the UK Rasch users' group, February 2008, Manchester, England.

Boursicot, K. and Roberts, T. (2006). Setting standards in a professional higher education course: Defining the concept of the minimally competent student in performance based assessment at the level of graduation from medical school, *Higher Education Quarterly, 60,* 74-90.

Bramley, T. (2005). A Rank-Ordering Method for Equating Tests by Expert Judgment. *Journal of Applied Measurement,* 6(2), 202-223.

Bramley, T. and Black, B. (2008). M*aintaining performance standards: aligning raw score scales on different tests via a latent trait created by rank-ordering examinees' work*. Paper presented at the Third International Rasch Measurement conference, University of Western Australia, Perth, January 2008.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17,* 59-88.

Fisher-Hoch H., Hughes S. and Bramley T. (1997). *What makes GCSE examination questions difficult? Outcomes of manipulating difficulty of GCSE questions.* Paper presented at the British Educational Research Association Annual Conference, September 1997, University of York.

Gill, T. and Bramley, T. (2008). *How accurate are examiners' judgments of script quality?* An investigation of absolute and relative judgements in two units, one with a wide and one with a narrow 'zone of uncertainty'. Paper presented at the British Educational Research Association Annual Conference, September 2008, Heriot-Watt University, Edinburgh.

Gill, T., Bramley, T. and Black, B. (2007). An investigation of standard maintaining in GCSE English using a rank-ordering method. Paper presented at the British Educational Research Association Annual Conference, September 2007, Institute of Education, London.

Idle, S. (2008). *An investigation of the use of the Angoff procedure for boundary setting in multiple choice tests in vocational qualifications.* A paper presented to the 34th annual conference of the International Association for Educational Assessment, Cambridge, UK, September 2008.

Impara, J. C. & Plake B. S. (1998). Teachers' ability to estimate item difficulty: a test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35(1),* 69-81.

Kimbell, R., Wheeler, T., Miller, S. and Pollitt, A. (2007). *E-scape portfolio assessment phase 2 report.* Goldsmiths, University of London.

Kolen, M. J. and Brennan, R. L. (1995). *Test Equating: Methods and Practices*. New York, NY: Springer Verlag.

Laming, D. (2004). *Human judgment.* London: Thomson.

Linacre, J.M. (2005). *Facets Rasch measurement computer program*. (Chicago, Winsteps.com).

Linacre, J. M. (2006). Rasch analysis of rank-ordered data. *Journal of Applied Measurement, 7(1),* 129-139.

Lorge, I. and Kruglov, L. (1952). A suggested technique for the improvement of difficulty prediction of test items. *Educational and Psychological Measurement, 12,* 554-561.

Newton, P. (2000). Maintaining Standards Over Time in National Curriculum English and Science Tests at Key Stage Two. *A report for the Qualifications and Curriculum Authority.*

Pollitt, A., Ahmed, A., & Crisp, V. (2007). The demands of exam syllabuses and question papers. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.)*Techniques for monitoring the comparability of examination standards.* London: Qualifications and Curriculum Authority. 166-206.

Raikes, N., Scorey, S. and Shiell, H. (2008). *Grading examinations using expert judgements from a diverse pool of judges.* A paper presented to the 34th annual conference of the International Association for Educational Assessment, Cambridge, UK, September 2008.

Reckase, M. D. (2006) Rejoinder: Evaluating Standard Setting Methods Using Error Models Proposed by Schulz. *Educational Measurement: Issues and Practice, 25(3*), 14-17.

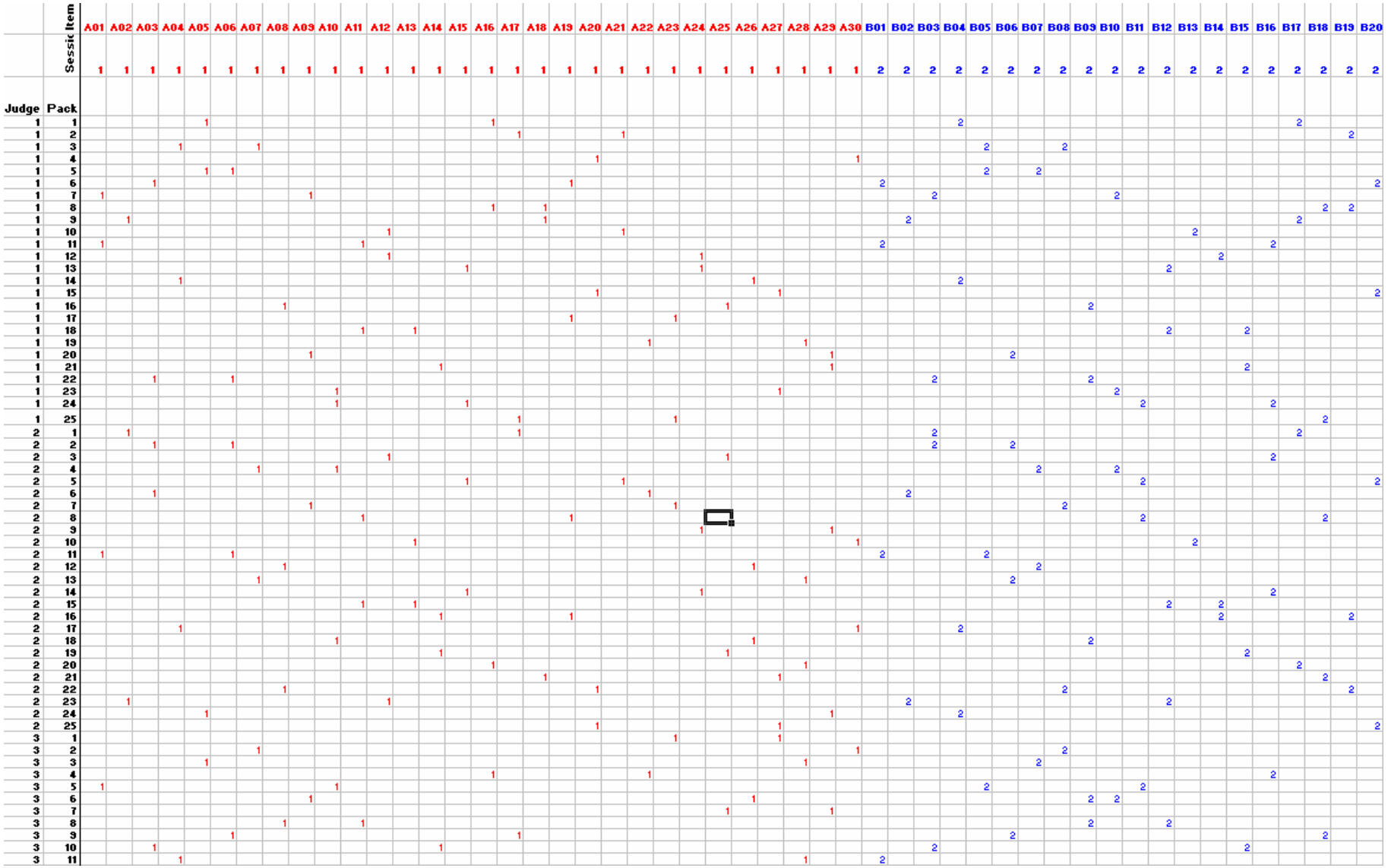
Schulz, E. M. (2006). Commentary: A response to Reckase's conceptual framework and examples for evaluating standard setting methods. *Educational Measurement: Issues and Practice, 25(3),* 4–13.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review, 3*, 273-286.

Thurstone, L. L. (1931). Rank order as a psychophysical method. *Journal of Experimental Psychology, 14,* 187-201. Chapter 10 in Thurstone, L.L. (1959). *The measurement of values.* University of Chicago Press, Chicago, Illinois.

Wright, B. D. and Masters, G. N. (1982). *Rating Scale Analysis.* Chicago: Mesa Press.

Appendix A: Stage 2 pack design showing judges, questions and packs



Judge	Pack	Item	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	A21	A22	A23	A24	A25	A26	A27	A28	A29	A30	B01	B02	B03	B04	B05	B06	B07	B08	B09	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20
		Session	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1	1						1											1																2																2			
1	2																		1				1																													2		
1	3					1			1																									2				2													
1	4																					1										1																					
1	5						1	1																										2				2		2											
1	6				1																1											2																				2
1	7	1									1																								2																		
1	8																	1		1																																2	2	
1	9		1																	1														2																2		
1	10													1									1																														
1	11	1											1																				2																				
1	12													1												1																											
1	13																1									1																										
1	14				1																							1								2														
1	15																		1										1																							2
1	16								1																		1																	2								
1	17																				1				1																											
1	18												1		1																														2			2					
1	19																							1						1																					
1	20										1																				1						2													
1	21															1															1																				
1	22				1			1																											2						2									
1	23											1								1																															
1	24											1					1																																		
1	25																		1							1																									
2	1		1																1																2																
2	2				1			1																											2					2									
2	3													1													1																	2						
2	4							1				1																															2			2					
2	5																1						1																												2
2	6				1																			1										2																
2	7										1														1																											
2	8												1									1																														
2	9																									1					1																				
2	10														1																	1																			
2	11	1						1																									2				2													
2	12								1																			1																							
2	13							1																						1								2												
2	14																1									1																									
2	15												1		1																													2		2					
2	16															1					1																									2					2	
2	17					1																										1				2													
2	18											1																1																						
2	19															1											1																							
2	20																	1												1																			
2	21																			1									1																				
2	22								1													1																												
2	23		1											1																				2																
2	24						1																								1				2													
2	25																					1							1																					2
3	1																								1				1																					
3	2							1																								1								2								
3	3						1																							1									2									
3	4																	1						1																									
3	5	1										1																						2									2						
3	6									1																	1															2	2						
3	7																										1				1																		
3	8								1				1																													2			2					
3	9							1											1																														2		
3	10				1											1																		2															
3	11					1																								1			2																

Appendix B: Instructions

March 2009

Instructions for completing the rank-ordering task on CPC Unit 1 (05598)

Please read carefully.

PACK CONTENTS

You have a set of 25 packs of CPC Unit 1 questions. Each pack contains 4 questions:

- Two questions from the current session (March 09)
- Two questions from the previous session (December 08)

WHAT TO DO

For each pack:

- Place the 4 questions into a single rank order from **most difficult (rank 1)** to **easiest (rank 4)**, based on ***your professional judgement*** of the relative difficulty of the questions.
- Using the question I.D. (e.g. A01) **record** your judgements on the recording form included in each pack. Please also answer the question about your confidence level on each recording sheet.
- When recording your judgements, please take extra care that you enter the rank-order into the table from **most difficult (1)** to **easiest (4)**.
- No tied ranks are allowed. If you are concerned that two or more questions are genuinely of exactly the same difficulty, indicate this by placing a bracket around them on the recording form. However, you must enter every question onto a separate line of the recording form.
- Return the four questions and the recording form to the pack. Start on the next pack.
- When you have completed all the packs, **record all your judgements once again** on the joint recording form enclosed.

HOW TO DO IT

This task is not an Angoff. Therefore, here are some guidelines on how to make your judgements:

✔ **Always** consider each question ***in comparison*** with the other questions in the pack.	***X*** **Do not** consider each question within a pack only on its own. Compare them with other questions in the pack.
✔ Comparing the 4 questions in the pack, think about which ones ***would*** be more difficult and which ones easier for candidates in general. You may find it helpful to focus upon a group of candidates that are familiar to you and how they would respond to the question.	***X*** **Do not** base your judgement of question difficulty on minimally competent candidates. Instead, think of candidates in general. If you find it helpful, focus on an actual group of candidates that are familiar to you.
✔ Use your ***professional judgement***: when comparing the 4 questions, take into account any features that, in your expert opinion, contribute to question difficulty.	***X*** There is no need to estimate the specific percentage of candidates that would get a question right. Just order the questions from most difficult to easiest in each pack.

FEEDBACK

When you have completed all of the packs, please complete the FEEDBACK QUESTIONNAIRE in as much detail as possible.

RETURNING THE MATERIALS

Finally, please return all the materials (questions in packs, completed recording forms, questionnaire, and expenses claim form for your postage expenses) in the envelopes provided using *NEXT DAY RECORDED DELIVERY*, **to arrive at Cambridge Assessment by Tuesday 17th March.**

Appendix C: FAQs

March 2009

FAQs about rank-ordering studies

What is the purpose of this research?
This is the second stage of the study that will investigate the use of a rank-ordering technique for determining the pass mark on tests containing multiple-choice questions for the purpose of standard maintaining from one session to the next. Rank-ordering all the questions on a single scale of difficulty should allow us to map a known pass mark from one test to an equivalent mark on the other test.

What should I do with the questions in each pack?
Your main task is to rank-order the four questions in each pack into a single rank order, from **most difficult (rank 1)** to **easiest (rank 4)** on the basis of question difficulty. Record your judgements on the individual recording forms and answer the question about your confidence level. Then return the questions and the recording form to the pack. Finally, record all your judgements once again on the joint recording form.

How should I arrive at a rank-order?
You should make a ***professional judgement*** of the difficulty of each question relative to other questions in a pack. As an expert in this qualification, you have probably developed an understanding of what makes questions difficult.

You may find it helpful to identify the most 'difficult' and the 'easiest' questions first, and then work out where the other two questions fit into the rank order.

Is this study a version of Angoff?
No. It differs in two important respects:

- You should make relative judgements and compare the questions in each pack ***with each other***, i.e. you should not give specific estimates for each question separately as you would in pre-Angoff. Always consider each question in comparison with the other questions in the pack.
- You should think of candidates in general and not just of minimally competent candidates. If you find it helpful, focus on an actual group of candidates you are familiar with.

Is there a 'right' answer to the order of the questions?
This is ***not*** a 'test'. The 'right' order of questions in any pack is the order that ***you*** determine by making a ***professional judgement*** about the ease or difficulty of each question relative to the others in the pack.

Are tied ranks allowed?
No tied ranks are allowed. If you are concerned that two or more questions are genuinely of exactly the same difficulty, you may indicate this by placing a bracket around them on the recording form, but you must enter every question onto a separate line of the recording form.

Are questions in packs arranged in any particular order?
You should make no assumptions about the way in which the questions from the different sessions are ordered within each pack. They are entirely random.

Can I collaborate with other colleagues who are completing the exercise?
Please do not collaborate with any of your colleagues who are completing this exercise as it is important that we have independent responses to the tasks. Besides, each of you will have packs containing different combinations of questions.

Should I complete the whole task in one go?
You can work flexibly to fit around your other plans and commitments. There is no need to complete the whole task in one sitting. However, we envisage that it should not take you longer than half a day to complete the task and the feedback questionnaire.

How long should each pack take me?
Gradually as you become accustomed to this task you will no doubt speed up. We envisage that it should take you about 5-10 minutes per pack.

What should I do with the materials?
Please return them in the enclosed envelope using *NEXT DAY RECORDED DELIVERY* **to arrive at Cambridge Assessment by 17th March 2009.**

What should I do if I have any other questions?
Please email (curcin.m@cambridgeassessment.org.uk) or phone me (01223 558774 - if necessary leave a message and I will phone you back).