Towards a suitable method for standard-maintaining in multiple-choice tests: capturing expert judgement of test difficulty through rank-ordering

Milja Curcin, Beth Black and Tom Bramley

Paper presented at the Association for Educational Assessment-Europe annual conference, Oslo, Norway, November 2010.

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

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

Introduction

In this paper, we report on a study that trialled and evaluated a rank-ordering method for capturing expert judgement of test difficulty for standard-maintaining on multiple-choice (MC) tests. The method is intended for use in situations when the requirements for statistical equating and linking of tests cannot be met, that is, if there are no common items or common persons between test sessions and no item pre-testing (see Kolen and Brennan, 2004). In this situation, expert judgement of test difficulty is necessarily called upon.

The method was trialled on multiple-choice (MC) units of two A-level[1] qualifications (one OCR and one CIE)[2] and two OCR vocational qualifications. Currently, a combination of expert judgement and statistical indicators of cohort attainment changes is used for setting pass marks/grade boundaries for these units in each session. The vocational units use the Angoff method (Angoff, 1971) for this purpose, while a method called awarding (see The Cambridge Approach, 2009) is used in the A-level units.

Though expert judgement of item difficulty or script[3] quality is taken into account in these methods, the issue of changes in test difficulty is generally approached indirectly via conclusions about changes or absence thereof in cohort attainment between sessions. These conclusions are reached based on various pieces of evidence, often most importantly statistical information reflecting changes in attainment. Due to unclear weighting of the contribution of expert judgement vs. statistical evidence, over-reliance on statistical evidence is often present in these two methods. This could lead to unnecessary grade boundary/pass mark changes and drift in standards over time even though the tests in consecutive sessions may have remained at a similar difficulty level. Indeed, there is evidence that experts unduly rely on performance data when these are available, at the expense of their independent expert judgement of test content difficulty (Clauser et al., 2009).

The rank-ordering method also relies on expert judgement, but it approaches the issue of test difficulty changes over time directly, rather than via conclusions regarding cohort attainment changes. This is important as the focus on measuring test difficulty can be seen as a pre-requisite of standard-maintaining, irrespective of any changes in cohort profile/ability over time. Trialling the rank-ordering method on the abovementioned tests allowed us to make some indirect comparisons of the three standard-setting/maintaining methods' respective assumptions, procedures and outcomes.

We make a distinction between standard-setting and standard-maintaining since the purpose and requirements of their respective procedures are different. Standard-setting is necessary when a new qualification is established in order to determine appropriate performance standards in relation to pre-determined content standards. This should enable fair distinctions between, for example, competent and not-yet-competent students ('pass/fail' distinction), or between different performance levels (e.g. grades A, B, C, etc.). Setting performance standards requires reference to expert judgement and values, and is a complex task (see e.g. Cizek, 2001; Cizek, Bunch, and Koons, 2004; Baird et al. 2000; Cresswell, 1996; Hambleton, 2001; Hambleton and Pitoniak, 2006; Newton, 1997, 2000; etc.).

The (iterative) Angoff method is a widely used standard-setting method. According to its procedures, in each new session, the expert judges initially individually estimate item

[1] A-level qualifications are usually studied over last two years of secondary school and are the standard entry qualification for academic courses in UK universities.

[2] OCR (Oxford, Cambridge and RSA) is one of the UK's leading providers of qualifications to learners of all ages while CIE (Cambridge International Education) is the world's largest provider of international qualifications for 14-19 year olds. They are both exam boards of The Cambridge Assessment Group, the largest assessment agency in Europe.

[3] A script is the examinee's responses to all the items in a test.

difficulty for minimally competent candidates (MCCs) in terms of the likelihood of their getting each item on a test right. Minimally competent candidates are usually defined as those candidates with sufficient skills to only just achieve a pass. At this initial stage, the judges do not have access to statistical performance data (e.g. facilities[4]) from the relevant session. Following this, the judges familiarise themselves with this data and participate in a meeting where they are allowed to discuss and alter their initial judgement. The average of all final estimates is then calculated, giving the recommended pass mark for the test. Finally, before the pass mark is determined, other statistical indicators, such as impact information (i.e. the proportion of students passing depending on the pass mark that is decided upon), are often considered. The final outcome may not actually be the one based either on the judges' initial or final judgements.

In place of an explicit mechanism for comparing test difficulty from different sessions, 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 (indirectly) as a standard-*maintaining* method, despite research evidence suggesting that an MCC's performance is difficult to conceptualise (e.g. Impara and Plake, 1998; Bouriscot and Roberts, 2006). Thus, the Angoff method lacks an explicit and direct mechanism for comparing tests from consecutive sessions. Nevertheless, it is often used in situations when a standard-maintaining method would be more appropriate.

Standard-maintaining methods are appropriate once a performance standard has been set for the first time for a particular test. Standard-maintaining session on session needs to take into account a multitude of factors that could lead to a drift in standards over time. Arguably, insistence on standard-maintaining over long periods may be neither feasible nor justifiable, given sometimes contextual changes such as (radical) curriculum changes, changes in educational policies, teaching standards, assessment purposes, values of the society, etc. (see Newton, 1997). However, in the shorter-term, especially where there have not been significant curriculum changes, one might reasonably expect consistent standard-maintaining to be achievable, enabling students of similar ability to obtain equivalent grades irrespective of current session's test difficulty (Newton, 2000).

The method that is currently used for standard-maintaining in the UK general qualifications (including A-level) is known as awarding. Its main aim is to determine key grade boundaries (typically A and E at A-level) for all subject components/units reflecting a performance standard that is consistent over time and comparable across components/units and specifications. Similarly to the Angoff method, this method is partly based on consideration of statistical performance data. Its judgemental aspect consists in experts' initially reviewing a selection of archive scripts that were awarded A or E in a prior session. This should enable them to "internalise" the performance standard corresponding to each grade boundary. They should then be able to consider the quality of scripts from the current session (several marks around each grade boundary) relative to the internalised performance standard and determine which of these scripts is the closest to this standard. They are also expected to take into account any changes in test difficulty between sessions. Consideration is also sometimes given to the Principal Examiner's personal judgement of the demand of the test.

It is apparent that the judgemental aspect of awarding demands a lot from experts in terms of internalising performance standards while there is no tangible evidence that they are indeed able to do this effectively. Furthermore, awarding lacks an explicit method for considering changes in test difficulty. Therefore, similarly to the Angoff method, awarding lacks a direct mechanism for comparing both performance and difficulty standards from one session to the

[4] For MC tests, facility is calculated as the proportion of students who answered a particular item correctly, ranging from 0 to 1 and it represents an indication of item difficulty for a given cohort, with higher values denoting easier items.

next. It should be noted that comparison with archive scripts is not done in awarding of MC tests as such “scripts” with their response strings of As, Bs, Cs, etc. do not support holistic judgement of quality.

As can be seen, both methods incorporate into their procedures a complex mixture of statistical performance indicators and expert judgement of either test difficulty or script quality, or a bit of both. Therefore, there is always a potential for tension between statistics and expert judgement, particularly if the relevant approach does not have an explicit mechanism of separating and weighting their respective influence on the final decisions regarding pass marks/grade boundaries.

This tension is not surprising since human (even expert) judgement is often seen as subjective, imprecise and unreliable. On the other hand, while statistics often offer the reassurance of being more objective, in the abovementioned situation they cannot be fully relied on since performance indicators for cohorts from different sessions cannot be assumed to mean the same thing. Therefore, they cannot directly link sessions in terms of test difficulty. Furthermore, the act of interpreting the various pieces of statistical information that contribute to a final decision in determining pass marks itself relies upon human judgement.

Nevertheless, currently in the context of the UK general qualifications at least, statistical information is often (implicitly) given more weight. Standards are increasingly maintained by reference to statistical performance indicators, without much influence of expert judgement of either test difficulty or script quality (Stewart, 2010; Grimston, 2010). Indeed, with respect to MC tests, in the current UK regulator’s Code of Practice (Ofqual, 2010) there is no explicit mention of the importance of establishing any changes in test difficulty independently of the cohort from one session to the next:

> *6.20 For units that are entirely composed of multiple choice items, the technical and statistical information specified in the ‘Quantitative’ section of paragraph 6.15 must be supplemented by item-level analyses, including facility and discrimination indices, and the correct answer to each item. The awarding committee must use a valid methodology to reach its grade boundary recommendations.* (Ofqual Code of Practice, 2010, p.44).

While “valid methodology” for this purpose is certainly necessary, it is unclear from this statement what is supposed to constitute it, how exactly the item-level analyses can be used to the end of standard-maintaining, and how prominent expert judgement of test difficulty should be in this process.

In such a situation, procedures that could clearly separate the judgemental from the statistical aspect of standard-maintaining would be useful. The rank-ordering method discussed in this paper could contribute to establishing such procedures as it keeps the judgemental strand of the standard-maintaining process separate from statistical information and other influences by capturing “pure” expert judgement. This is to say that the experts do not have access to statistical information while rank-ordering items in terms of difficulty. Unlike current standard-setting methods, rank-ordering also enables direct comparison of tests from two or more sessions, allowing “test equating” by expert judgement based on perceived test difficulty. In this way, it avoids having to resort to any “conceptualisations” of MCC ability or “internalisations” of performance standards. Thus, in theory, it has several advantages over methods currently used for similar purposes.

The next section contains a brief description of the rank-ordering method as it has been used so far. We then present the methodology of the current study, followed by a presentation of the most important results and a discussion.

Previous research on the use of rank-ordering as a standard-maintaining method

The current study employs a version of the rank-ordering method developed by Bramley (2005, cf. Thurstone, 1931) for standard-maintaining based on expert judgement of script quality. The method is an extension of the paired comparisons method for capturing relative judgements of non-physical attributes, e.g. ‘seriousness of crime’, allowing measurement of these attributes that is more akin to that of physical attributes such as temperature or weight (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 much of the trait it is judged to possess. Rank-ordering produces a similar outcome through a more efficient procedure since rank-ordering a set of, say, 10 scripts, yields the equivalent of 45 paired comparisons.

Several rank-ordering studies 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 (e.g. Black, 2008; Black and Bramley, 2008). In these studies, which used tests containing mainly open-ended items, the judges rank-ordered a number of scripts from different sessions within several packs from best to worst in terms of quality, making holistic judgements and taking account of their perception of any differences in difficulty between the pair of tests involved. They had no access to original marks. Based on these ranks, a measure of script quality is derived (see next section for details). This measure correlated very well (r=0.8 or 0.9) with the original marks, while the method has proven to be robust, rigorous and capable of being cross-validated (see Bramley and Gill (2010) and Black and Bramley (2008) for a more detailed evaluation).

However, as we already noted, script quality judgements are less appropriate for standard maintaining on objective tests. Indeed, even in the case of more subjective tests, measuring changes in test difficulty without reference to examinee ability would address the issue of standard maintaining more directly than anything else (Bramley, 2010). The question is whether this can be done reliably by consulting expert judgement of test difficulty.

The study reported here investigated this issue by adapting the rank-ordering method to be used for capturing judgements of item and test difficulty of MC tests in two vocational and two A-level qualifications. The findings regarding the vocational qualifications were reported in Curcin, Black and Bramley (2009) and are summarised here together with the results for the two general qualifications.

An additional strand of this study also looked at the use of ‘non-expert’ judges. Hitherto, rank-ordering for the purpose of standard maintaining has involved making relative judgements about script quality, for which expert judges are probably the most appropriate. However, the task here is to judge the difficulty of questions from the point of view of test-takers. As such, we wished to investigate the merits of students, i.e. those people who directly experience question difficulty, ranking items according to their own perception of question difficulty.

Method

Design, judges and procedures

In OCR vocational qualifications we used MC units from Certificate of Professional Competence in Road Haulage and Passenger Transport (CPC) and Award in Administration (AinA). These units are assessed by 30-item MC tests at the end of the course and either a pass or a fail can be achieved on each. The general qualifications used were OCR A-level Critical Thinking (CT) and CIE A-level Biology. Again, only their units assessed by MC tests (15 and 40-item respectively) were used in this study.

The judges recruited were subject experts who normally took part in the standard-setting/maintaining procedures for these qualifications, but also worked in other professional roles in respect of these qualifications – for instance, as examiners, teachers, item writers, trainers, etc. In addition to the expert judges, we recruited another pool of judges for a separate rank-ordering task for the Critical Thinking test – 194 students who were at the time preparing for the CT January 10 exam (see below for details).

For each test, the rank-ordering task was conducted in two stages, with largely the same expert judges participating both times. Table 1 illustrates the design of the study. It can be seen that for each qualification and stage the expert judges compared items from two consecutive tests. One test was always the same between stages (shaded boxes), which enabled us to investigate the consistency of the rank-ordering judgements (except for the student judges, who carried out the rank-ordering task only once). During each exercise, the judges had no access to statistical performance data.

Table 1: Design of the study

Type	Qualification	Session	Stage 1	Stage 2
Vocational	CPC	September 08	✓	
		December 08	✓	✓
		March 09		✓
	AinA	November 08	✓	
		April 09	✓	✓
		June 09		✓
A-level	Biology	November 08	✓	
		June 09	✓	✓
		November 09		✓
	CT - experts	January 09	✓	
		June 09	✓	✓
		January 10		✓
	CT - students	January 09	✓	
		June 09	✓	n/a

The expert judges carried out each rank-ordering task at home. In each stage, they were given 25 or 30 packs (depending on the qualification) containing three or four items from different sessions each (e.g. two from September 08 and two from December 08). The items were presented on individual sheets of paper to enable easier physical rank-ordering. The students carried out the task as part of their CT lessons. They were given four packs of three items each, presented in the same way as to the expert judges.

For each qualification and stage, Table 2 summarises the information about the number of judges, number of packs and items in packs.

Table 2: Judges, packs and items

Qualification		Judges		Packs	N items in packs
		Stage 1	Stage 2		
CPC		6	7	25	4
AinA		4	5	25	4
Biology		8	8	25	4
CT	Experts	8	8	15	3
				15	4
	Students	194	n/a	4	3

The pack design was generated by random allocation of items to packs, and was then "tweaked" to meet several criteria, most importantly: each judge's exposure to each item at least once; minimising the number of times a judge saw any one item (no judge saw any item more than twice); and, for each judge to have a unique set of packs and combination of items. Since the data were analysed by fitting a Rasch model, it was also necessary to ensure sufficient linking of all items involved in comparisons (see Linacre, 2005).

The judges were given detailed written instructions about how to carry out the rank-ordering task, and had no training or a practice session prior to the main task. The experts were instructed to rank-order the items in each pack from easiest to most difficult for a familiar or an average group of students or for students in the relevant qualification in general, and record their responses on a recording sheet. They were asked to use their professional judgement and were not given explicit guidelines regarding how to judge item difficulty. The students were asked to carry out essentially similar task to the experts, except that they were to rank-order items in terms of difficulty *for themselves,* as if they were seeing and doing the items for the first time.

Data analysis

The ranks obtained in each stage 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 item *i* beats item *j* in a paired comparison
and θ_i = the measure for item *i*
and θ_j = the measure for item *j*

The analysis produces a latent trait scale (i.e. a common scale of difficulty, ranging from negative values (easier) to positive values (more difficult)), and its unit, 'logit', denotes the amount of difficulty (i.e. measure) each item was perceived to have relative to the scale origin.

The next stage of the analysis was the test equating, which allowed us to determine the pass mark on the current test in relation to the pass mark on the previous test. Once the perceived difficulties of the items 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.

Test Characteristic Curves (TCC) are plots of expected score on a test against ability. For a dichotomous item, the expected score is given by the equation for the Rasch model:

$$ln[P_{ji} / (1-P_{ji})] = \theta_j - b_i \qquad (1)$$

where P is probability of success of person j on item i, while θ_j represents person ability and b_i represents item difficulty. The expected test score is the sum of the expected scores on each item for a candidate of a given ability:

$$TS_j = \sum_{i=1}^{N} P_i(\theta_j) \qquad (2)$$

where TS_j is the expected test score for examinees with ability level θ_j, i denotes an item and $P_i(\theta_j)$ is obtained via equation (1).

If the item 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 (see e.g. Wright and Stone, 1979, p64-5). The abilities corresponding to each possible raw score on the test[5] were obtained by this method. TCCs can then be plotted based on these results. If the TCCs for the two tests are plotted on the same graph then it is possible to find the pass mark on one test corresponding to a given pass mark on the previous test.

The next step was the evaluation of the rank-ordering judgements underlying the equating results. They can be evaluated by investigating the fit of the data to the Rasch model (item residuals[6]; separation and separation reliability (see Fisher, 1992)), judgement consistency between repeated tests, and the correlation between the measures of difficulty obtained in the rank-ordering exercise and empirical facilities (see below) for the relevant session.

The measure-facility correlations were calculated within test (for instance, for CT stage 1, they were calculated separately for January and June sessions). Although the purpose of the rank-ordering exercise was to provide a mechanism for comparing two or more sessions within each pack, the analyses were necessarily conducted within session since we could not assume that the cohorts between sessions were comparable. In other words, a facility of 0.75 in two different sessions may not necessarily indicate the same level of item difficulty. However, it is plausible to assume that the results within session are generalisable to those between sessions.

Importantly, as Bramley and Gill (2010) point out, the rank-ordering method is a ‘strong’ method in that it can be invalidated in two distinct ways. It is possible that (a) the scale of perceived difficulty does not agree with the empirical facilities, and (b) the data may fail to fit the model and/or fail to create a meaningful scale. If either of these problems is detected, this can call into question the validity of the final equating result.

[5] The abilities corresponding to the maximum score and the score of zero were extrapolated from the other values as it is impossible to estimate measures for extreme scores.

[6] Residuals lower than 2.5 were considered acceptable.

Results

In this section, we first present the expert judges' and then the CT students' results. In each case, we consider the following issues:

a. How well did the pass marks/grade boundaries produced by the rank-ordering method on its own agree with the official pass marks/grade boundaries produced by the Angoff and awarding methods?
b. How well did the rank-ordering data agree with the Rasch model?
c. How consistent were the expert judges' rank-orders of the same items on two occasions? (This is not considered for CT students as they did the rank-ordering task only once).
d. How well did the measures of perceived difficulty obtained in the rank-ordering exercises agree with empirical facilities?
e. Finally, how well did the measures of perceived difficulty obtained from CT experts agree with those obtained from CT students?

Expert judges

Equating results

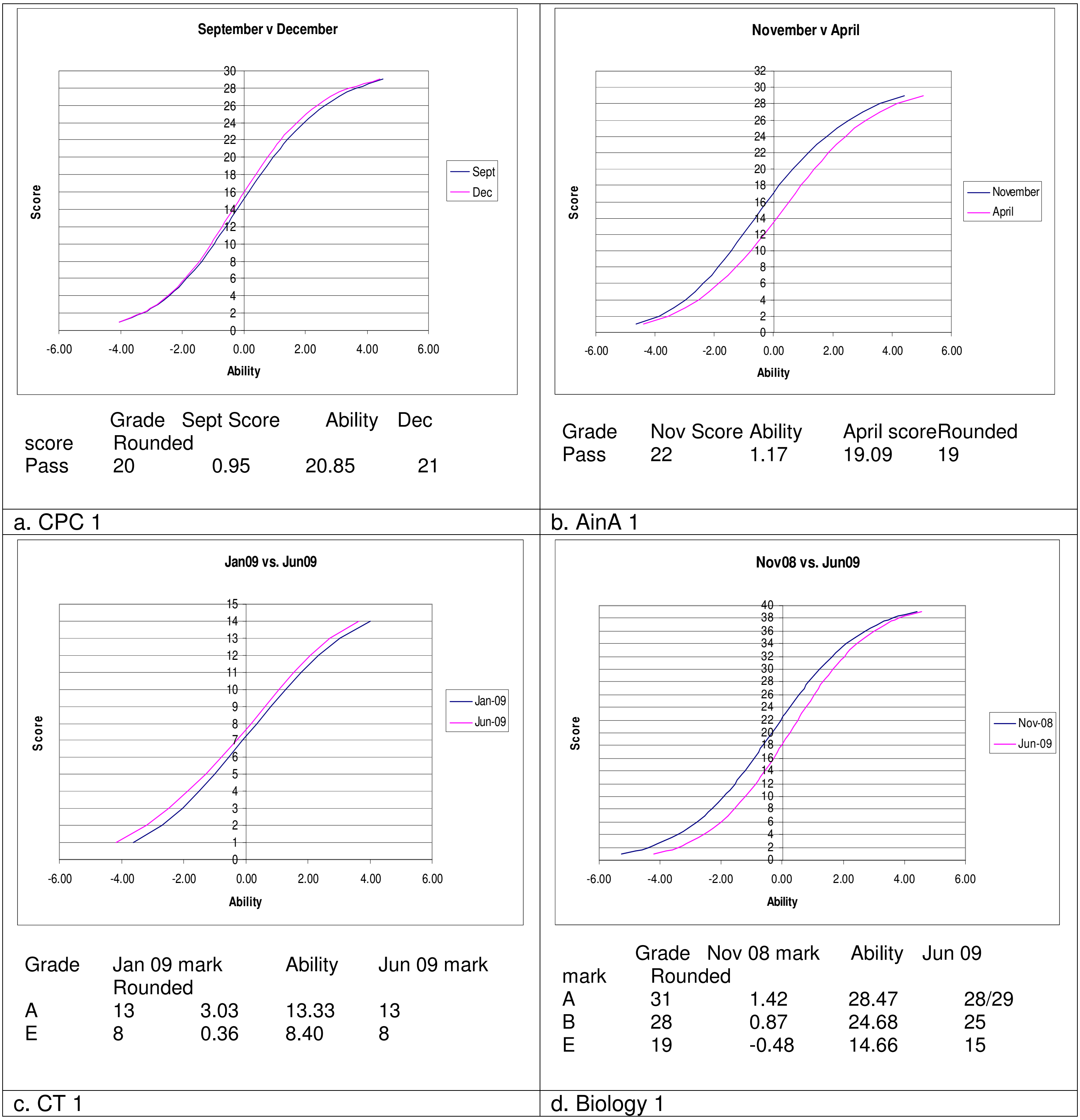
The equating procedure described in the previous section was used to produce the TCC graphs presented in Figure 1 on the next page. By way of example, we present four out of eight graphs based on expert judges' rank-orders.

If the two tests were of equivalent difficulty according to the judges, the same estimated ability level would correspond to the same cut score on both tests (i.e. the curves on the graphs would overlap). The tables below each graph show the official marks for each grade boundary for the earlier of the two sessions alongside the corresponding ability level (in logits) and the mark for each grade boundary for the following session based on the outcome of rank-ordering (together with the rounded value for this mark).[7]

According to expert rank-ordering judgements, the December test for CPC stage 1 was perceived as easier than the September test. In other words, students of equivalent ability would have achieved the score of 21 in December, and 20 in September. In AinA stage 1, the April test was judged as more difficult than the November test, that is, students of equivalent ability would have achieved a lower score (19 instead of 22).

For CT 1, the June 09 test was judged by the experts to be very close in difficulty to the preceding January 09 test. If the scores were rounded to the nearest whole number, which is necessary to make the scores usable in practice, the grade boundaries would have remained the same based on the rank-ordering recommendation alone. In Biology 1, the November 08 test was perceived to be easier than the June 09 test, though there was more of a difference at lower ability levels.

[7] Note that the key grade boundaries for CT are A and E, and for Biology A, B and E.



	Grade	Sept Score	Ability	Dec
score	Rounded			
Pass	20	0.95	20.85	21

a. CPC 1

Grade	Nov Score	Ability	April score	Rounded
Pass	22	1.17	19.09	19

b. AinA 1

Grade	Jan 09 mark		Ability	Jun 09 mark
	Rounded			
A	13	3.03	13.33	13
E	8	0.36	8.40	8

c. CT 1

	Grade	Nov 08 mark	Ability	Jun 09
mark	Rounded			
A	31	1.42	28.47	28/29
B	28	0.87	24.68	25
E	19	-0.48	14.66	15

d. Biology 1

Figure 1: Sample of equating outcomes

Tables 3 and 4 present pass marks and grade boundaries established through official Angoff and awarding procedures, compared (reading horizontally) with the rank-ordering recommendations for the corresponding sessions. The pass marks were generally comparable, though in a few cases they were quite different (e.g. CT January 10 E boundary, or Biology June 09 for all boundaries).

Table 3: Angoff and rank-ordering pass marks, compared with the endorsed pass marks (from Curcin et al., 2009)

Qualification	Session	Endorsed	Angoff final	Rank-ordering
CPC	Sept 08	20	21	n/a
	Dec 08	21	21	21
	Mar 09	21	21	21
AinA	Nov 08	22	21	n/a
	April 09	20	21	19
	June 09	21	20	19

Table 4: Awarding and rank-ordering grade boundaries

		Awarding			Rank-ordering		
Qualification	Session	A	B	E	A	B	E
CT	Jan 09	13	n/a	8	n/a	n/a	n/a
	June 09	13	n/a	8	13	n/a	8
	Jan 10	13	n/a	8	12	n/a	5
Biology	Nov 08	31	28	19	n/a	n/a	n/a
	June 09	31	28	18	28/29	25	15
	Nov 09	30	27	18	32	29	20

Although these sorts of outcome comparisons give us some indication as to how useful/correct the rank-ordering judgements may have been in determining pass marks/grade boundaries, it is difficult to establish whether the rank-ordering or the official outcome is the “correct” one. Furthermore, a (direct) comparison of different outcomes is problematic since the Angoff method and awarding involve a mixture of statistical information and expert judgement, while the rank-ordering outcome is based *solely* on expert judgements of test difficulty. The most one could hope for in this situation is for these approaches to inform each other prior to reaching a final decision rather than compete in outcome “correctness”. Ultimately, there is currently no “perfect” standard-maintaining method that we could use to compare the rank-ordering results against. If there were, we would not be exploring the alternatives.

Reliability and fit to the Rasch model

The rank-ordering data fit the model well in all stages of the study, with item residuals largely within acceptable limits. In addition, item separation ranged from 2.61 to 5.26 and separation reliability from 0.87 to 0.97.[8] In the current context, high separation reliability would suggest that the judges perceived a similar scale of difficulty for the items in question, i.e. that the differences among difficulty measures did not arise from random variation.

[8] Note that these separation reliability coefficients are likely to be overestimates because of violation of local independence in the rank-ordering method (Linacre, 2006). Separation reliability less than 0.5 implies that the differences between measures are mainly due to measurement error (see e.g. Fisher, 1992). According to Linacre (2009), separation reliability of 0.8 is the lowest reliability for serious decision-making.

Table 5: Separation and separation reliability coefficients

Qualification	Stage	Average N judgements per item	Separation	Separation reliability
CPC	1	30	2.82	.89
	2	35	2.72	.88
AinA	1	24	2.75	.88
	2	30	2.75	.88
Biology	1	30	2.64	.87
	2	30	2.61	.87
CT	1	72	5.26	.97
	2	72	4.65	.96

CT judges were able to make more judgements per item (see Table 5) in the same amount of time as the other judges because the CT test had fewer items. This is likely to have at least partly contributed to the higher separation and separation reliability coefficients for CT; though these may also possibly be partly explained by a higher reliability of judgements (i.e. greater agreement among judges on item difficulty rank-orders). Overall, though, all the abovementioned reliability indices are at a satisfactory level to indicate that the rank-ordering method produced a meaningful scale of item difficulty.

Judgement consistency

Another way of evaluating the reliability of the rank-ordering method and judgements is by investigating how repeatable the judgements were when elicited for the same items on two occasions. On each occasion, these items were combined with a different set of items, so that item combinations were unique each time. A correlation between the measures of difficulty obtained for the same items on two occasions would give us a measure of judgement consistency. A poor relationship would suggest that the judges did not have a clear, consistent view of each item's relative difficulty.

Table 6 shows that the correlations for all sets of judges were very good, though somewhat lower for the two vocational qualifications.

Table 6: Judgement consistency

Qualification	Session	Pearson correlation
CPC	Dec 08 1 vs. Dec 08 2	.656
AinA	April 09 vs. April 09 2	.658
Biology	Jun 09 1 vs. Jun 09 2	.773
CT	Jun 09 1 vs. Jun 09 2	.952

This provided evidence that the rank-ordering judgements elicited in this study were consistent and reflected judges' actual perception of item difficulty to a great extent. This further demonstrates the reliability of the rank-ordering method when used in this context.

Measure-facility agreement

A within-session correlation between the rank-ordering measures of relative item difficulty and the empirical facilities for corresponding items gave an indication of the extent to which the respective rank-orders of these values were in agreement, that is, how "correct" the rank-ordering judgements of our judges were. A strong negative correlation would suggest a high level of agreement, i.e. that the lower facility values correspond to higher measures. This in turn would mean that the judges were good at judging relative item difficulty for students, giving us confidence that our equating results were valid and based on sound expert judgement.

Table 7 lists Spearman correlation coefficients for each test and stage of the rank-ordering exercise.[9] All correlations were low to moderate. The CPC correlations were generally the lowest, while the CT ones were the highest. It is also apparent that there was some variability in correlation size between each pair of sessions, particularly in the case of Biology and AinA.

Table 7: Rank-ordering measure-facility correlations

Qualification	Stage	Session	Spearman correlations
CPC	1	Sept 08	-.260
		Dec 08	.230
	2	Dec 08	-.064
		Mar 09	-.178
AinA	1	Nov 08	-.459
		April 09	-.337
	2	April 09	-.343
		Jun 09	-.629
Biology	1	Nov 08	-.523
		Jun 09	-.368
	2	Jun 09	-.358
		Nov 09	-.160
CT	1	Jan 09	-.604
		Jun 09	-.668
	2	Jun 09	-.572
		Jan 10	-.471

By way of evaluation of the rank-ordering results for the vocational tests, these correlations can be compared with the correlations between average *initial* Angoff difficulty estimates and empirical facilities for the corresponding sessions, presented in Table 8. These correlations are comparable in size to all rank-ordering correlations except those for CPC, which were significantly higher. It is unclear why the CPC correlations were that low in the rank-ordering method, and a replications study could be carried out to try and determine this.

Table 8: Correlations between empirical facilities and initial Angoff estimates (from Curcin et al., 2009)

Qualification	Angoff session	Spearman correlation
CPC	Sept 08	.613
	Dec 08	.358
	Mar 09	.301
AinA	Nov 08	.615
	April 09	.534
	June 09	.612

It is clear that the rank-ordering method failed to elicit judgements that were generally better correlated with the facilities than the initial Angoff estimates. However, one notable difference

[9] Spearman rank correlation was used as the relationship between facility values and measure is potentially non-linear, because the facility values can suffer from floor and ceiling effects.

between these two approaches is that the Angoff method *relies* on the difficulty estimates being as accurate as possible (which according to many studies appears to be an unattainable goal, e.g. Brandon, 2004; Thorndike, 1980; Morrison et al. 1994; Impara and Plake 1998; Goodwin, 1999; Idle, 2008), while the rank-ordering method does not require this accuracy, recognising in this way the inherently variable and subjective nature of expert judgement in this domain.

Importantly, none of the current rank-ordering correlations were as high as those obtained in most previous rank-ordering exercises where judges were rank-ordering scripts, rather than items, in terms of quality. This is perhaps unsurprising considering the widely held view that item difficulty is difficult to judge, particularly if it has to be judged for somebody else (in our case, judges judging item difficulty for students).

In the rank-ordering context, the challenge of judging item difficulty for somebody else can be seen as a problem in the domain of the trait actually being measured when we try to measure test difficulty. While expert judges might agree among one another as to the relative order of item difficulty on a test, the question is how this rank-order relates to student performance on that test. For instance, in CT 2, there appeared to be a high level of agreement between judges, judging by high separation reliability of 0.96. At the same time, the corresponding measure-facility correlations were only moderate at -.572 and -.471 for June 09 and January 10 respectively.

For those sessions where the correlations were at least moderate we can perhaps be reassured that the test equating results based on the corresponding rank-ordering judgements were most likely plausible and could be taken as a useful indication of changes in test difficulty from one session to the next.

CT students as judges

Equating results

Recall that the grading outcome based on CT expert judges' rank-ordering results indicated that there was no major difference in test difficulty between Jan 09 and Jun 09 sessions (Figure 1). Applying the same equating method in the case of CT students, we obtained a similar outcome based on their judgements (Figure 2) for the same pair of tests. This finding is reassuring, providing further support for the grade boundary recommendation based on CT expert judgements.

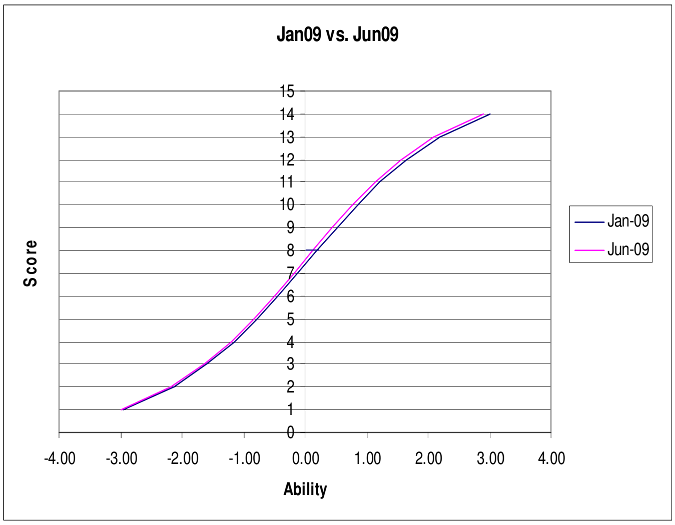


Figure 2: Equating outcomes based on the rank-ordering judgements of CT students

Reliability and fit to the Rasch model

The FACETS analyses showed that both judge and item data fit the Rasch model well. The separation indices of student judgements are comparable with those of experts though somewhat lower. This suggests that there was somewhat less agreement between student judges than between expert judges regarding the perceived rank-order of question difficulty in these two sessions.

Table 9: Separation and separation reliability coefficients for student judges

Average N judgements per item	Separation	Separation reliability
157	3.76	.93

Measure-facility agreement

Table 10 shows the measure-facility correlations obtained based on student judgements. It can be seen that the correlations were moderate, and comparable to those of expert judges for the corresponding sessions, though the experts' ones were a bit higher. This suggests that the students and experts had a similar view of the difficulty of the two tests.

Table 10: Measure-facility correlations

Session	Spearman correlations
Jan 09	-.549
Jun 09	-.559

Indeed, a comparison of difficulty measures obtained from the students' and experts' rank-ordering judgements produces very high correlations (r=.827 for Jan 09 and r=.887 for Jun 09).

Finally, it should be noted that the CT students' separation coefficients were lower than those for the CT experts (despite a higher number of judgements per item). This suggests that there is no reason to support the view that students are better than experts in the role of judging relative item difficulty (from this data at least). However, there may have been contextual or motivational factors that may have impacted upon the reliability of judgements for these participants (e.g. the ranking was done in class, rather than at home; despite instructions to the contrary, not all students had taken this exam; unlike the experts, they were not financially rewarded for the task and nor did they have any enduring professional relationship with Cambridge Assessment to maintain).

Discussion

In this study we have presented a way of equating tests from different sessions based on expert and student judgements of relative item difficulty captured through rank-ordering. The equating results obtained generally produced comparable grade boundary/pass mark recommendations to those established in the official awarding/standard-setting procedures. This is encouraging considering that the rank-ordering results were based solely on test difficulty judgements, without access to statistical information or consideration of impact data.

One issue with the current approach is the possibility of artefacts related to the equating procedures used. As long as the resulting distributions of perceived item difficulty are the same as the distributions of empirical difficulty (e.g. have the same mean and standard deviation) the same equating relationship will obtain even if the perceived difficulty values individually bear no relationship at all to the empirical values. This is why it is necessary to investigate the relationship between perceived and empirical difficulty to validate the method. A similar situation can occur in the Angoff method – the average of the judges' estimated probabilities can converge on the "correct" cut-score even if there is no relationship between the estimated probabilities and empirical data at the individual item level.

The equating results we obtained were based on quite reliable and consistent rank-ordering judgements (particularly in the case of CT). In addition, there was little difference between CT students' and CT judges' measure-facility correlations, as well as a high correlation between the students' and the judges' rank-orders for the same sessions. Taken together, these findings suggest that the rank-ordering method has no problem eliciting consistent and reliable rank-ordering judgements capturing something of the judges' genuine view of item difficulty irrespective of the qualification type.

However, consistency does not always guarantee judgement "correctness" in the sense in which it has been defined in this study. Indeed, we have seen some fairly low measure-facility correlations in all the subjects investigated so far though the majority of the correlations were moderate. We have also observed variability in judgement correctness between qualifications. Two possible explanations for this variability in terms of professional experience or nature of the subject are considered below.

With respect to differences in judges' professional experience, for instance, all CT judges were currently or recently involved in teaching, while Biology judges were less so, but had more MC item-writing experience. This possibly suggests that contact with students might be more important than item-writing experience for this particular purpose. The difference in correlations between vocational and A-level tests might be because the CPC and AinA judges had less interaction with the students on a regular basis than either the CT or Biology judges due to the nature of these qualifications. A more controlled study investigating the influence of these aspects of professional expertise on judging item difficulty might be the next step in getting to the bottom of this issue.

Another aspect that might distinguish CT (with the highest level of correctness in rank-ordering judgements) from the other three tests is to do with the nature of CT as a subject. CT is more skills-based than the other subjects discussed here, closely intertwining factual knowledge with the relevant skills that need to be demonstrated in the response to every item on the test. Thus, unlike in the other three subjects, no CT item tests factual knowledge on its own. As skills could be said to form a more natural hierarchy than factual content, it is possible that the CT judges benefited from this, finding it easier to rank-order items in terms of skill they test rather than having to consider the likelihood of students' possession of a certain piece of factual knowledge. Their questionnaire responses indicated that they did perceive a fairly straightforward and similar rank-order of CT skills from the basic to the more advanced ones. In contrast, in other tests, the judges tended to separate factual knowledge from the skills that may have been required to respond to an item correctly. Therefore, they may have found it more difficult to perceive a similar rank-order in the likelihood of student knowledge of certain facts, possibly because these are generally not acquired in a predictable, orderly way, as much as skills (at least in some domains) can be.

Whatever the reason for between-subject differences in correlations, all the measure-facility correlations in this study were generally lower than the measure-*mark* correlations observed in previous rank-ordering studies. This may be related to the fact that judging item difficulty is perhaps a more complex (or less familiar) task than judging script quality. Various studies suggest that it is not easy to conceptualise in any precise and straightforward way what actually constitutes or contributes to an item's difficulty (see e.g. Pollitt, Ahmed and Crisp, 2007; Pollitt, Hutchinson, Entwhistle, and de Luca, 1985; Nathan and Koedinger, 2000; Gorin and Embertson, 2006). Judging by the example of CT 2, it is even possible for judges to agree to a great extent regarding relative item difficulty on a test, but for their judgements to still be mismatched with actual student performance (i.e. empirical facilities).

Interestingly, the correlation between CT judges' difficulty judgements and the difficulty *judgements* (rather than actual *performance* on items) of CT students was very high. This might be because both sets of difficulty measures were based on two sets of judgements (rather than judgement vs. performance) capturing a similar trait, i.e. perceived item difficulty. The situation in script quality rank-ordering studies is partly similar to this. Both measures of script quality and marks are based on two (similar) sets of judgements, capturing a similar trait, i.e. perceived script quality.

Therefore, we might expect less of a mismatch between two similar sets of judgements about a similar trait (be it a comparison of script quality judgements with marks, or expert and student judgements of item difficulty) than between judgement and performance measures (difficulty measures vs. empirical facilities). The latter mismatch may be putting a limit on the size of correlations that can reasonably be expected in this sort of exercise if we are using empirical facilities as a point of comparison, in addition to any other problems that might exist (e.g. disagreement between judges). Perhaps this needs to be taken into consideration when evaluating the results of any item difficulty rank-ordering exercise in terms of measure-facility correlations.

It is difficult to say at this point what might be the upper limit of measure-facility correlations in this context and further research and/or practice is needed to establish this. However, exceptionally low correlations we might consider to be a questionable basis for any equating procedures and pass mark decisions. They require further investigation to establish what has caused them to be so low. For instance, was this maybe related to judge expertise, nature of the subject, or lack of training and understanding of the rank-ordering process? Of course, alongside measure-facility correlations, another important consideration is agreement between judges, as evidenced in Rasch reliability statistics, as well as difficulty judgement consistency/repeatability. High reliability levels (as in the case of CT) give more credibility to the rank-order judgements and grade boundary recommendations based on them even though they may be accompanied by fairly low measure-facility correlations.

We have seen that the item difficulty judgements elicited by the rank-ordering method are not necessarily much more correct than those elicited by other methods. However, given its other advantages, there is arguably the case for pursuing the investigations and possible use of this method.

As Clauser et al. (2009) observe, there is currently no available method that provides accurate content-based judgements of item difficulty independently of performance data and therefore methods such as Angoff, though they have their problems, cannot be abandoned.

However, we should perhaps consider methods that do not require "accurate" but rather approximate content-based judgements that are independent of performance data, rather than trying to increase the accuracy of expert judgement by exposing experts to this data. It seems that approximation is the most we could aim for in the domain of expert judgement of item difficulty in any case.

In general, since rank-ordering judgements of test difficulty are not influenced by statistical information, group dynamics, impact information, etc., they can be independently evaluated in a number of ways (judgement accuracy, consistency and reliability), and the rank-ordering recommendations can be taken into consideration or not depending on the evaluation outcome. The rank-ordering method and recommendations could therefore be incorporated into current standard-maintaining processes as an independent piece of evidence, alongside statistical performance indicators, thus contributing to face validity and defensibility of these processes.

Although it does not always provide "accurate" difficulty judgements, rank-ordering has several advantages over comparable judgemental methods for judging item/test difficulty. It does not require accurate judgments in the first place (unlike the Angoff method), recognising in this way the inherently imprecise and subjective nature of expert judgement in this domain. Importantly, also, the method allows direct comparison of tests from two or more sessions (and the comparison is based on test difficulty rather than script quality); it does not require conceptualisation of any particular competence level or performance standard; it eliminates internal standards of the judges; and it separates the judgemental process from other sources of evidence such as statistical information (though it does not preclude the use of statistical information as an independent source of evidence if necessary).

References

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement, 2,* 449-460.

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.

Baird, J. (2000). Are examination standards all in the head? Experiments with examiners' judgements of standards in A-level examinations. *Research in Education*, 64, 91-100.

Black, B. (2008). *Using an adapted rank-ordering method to investigate January versus June awarding standards*. A paper presented at the Fourth Biennial EARLI/Northumbria Assessment Conference, Berlin, Germany, August 2008.

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

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. (2010). *Locating objects on a latent trait using Rasch analysis of experts' judgments.* Paper presented at the conference "Probabilistic models for measurement in education, psychology, social science and health", Copenhagen, Denmark, June 2010.

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

Bramley, T., & Gill, T. (2010). Evaluating the rank-ordering method for standard maintaining. *Research Papers in Education,* 25(3), 293-317.

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

Cambridge Assessment (2009). The Cambridge Approach: Principles for designing, administering and evaluating assessment. Cambridge: A Cambridge Assessment Publication.

Cizek, G. J. (Ed.) (2001). *Setting performance standards: Concepts, Methods and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek, G.J., M.B. Bunch, and H. Koons (2004). An NCME instructional module on setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice,* 23(3), 31–50.

Clauser, E. B, Mee, J., Baldwin, S. G., Margolis, M. J. and Dillon, G. F. (2009). Judges' Use of Examination Performance Data in and Angoff Standard-Setting Exercise for a Medical Licensing Examination: An Experimental Study. *Journal of Educational Measurement*, 46 (4), 390-407.

Cresswell, M.J. (1996). 'Defining, setting and maintaining standards in curriculum-embedded examinations: judgmental and statistical approaches.' In: Goldstein, H. and Lewis, T. (Eds), *Assessment: Problems, Developments and Statistical Issues*. Chichester: John Wiley & Sons.

Curcin, M., Black, B. and Bramley, T. (2009). *Standard maintaining by expert judgment on multiple choice tests: a new use for the rank-ordering method*. Paper presented at the British Educational Research Association conference, University of Manchester.

Fisher, W.P.J. (1992). Reliability statistics. *Rasch Measurement Transactions*, 6(3), 238

Goodwin, L. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. *Applied measurement in Education, 12, 13-28.*

Gorin, J. S. and Embretson, S. E. (2006). Item Diffficulty Modeling of Paragraph Comprehension Items. *Applied Psychological Measurement,* 30, 394-411.

Grimston, J. (2010). Secret downgrading of GCSE exam results. In Sunday Times 28/02/10. Available online at: http://www.timesonline.co.uk/tol/news/uk/education/article7043962.ece Accessed 23/07/10.

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

Hambleton, R. K. and Pitoniak, M. J. (2006). Setting Performance Standards. In Brennan, R. L. (Ed.) *Educational Measurement.* American Council of Education, Praeger, 433-470.

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.

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.

Linacre, J. M. (2009). Practical Rasch Measurement Course, Lesson 3. Available through www.statistics.com.

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

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

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

Morrison, H.G., Busch, J.C. and D'Arcy, J. (1994). Setting reliable National Curriculum standards: a guide to the Angoff procedure. *Assessment in Education, 1(2),* 181-99.

Nathan, M. J. and Koedinger, K. R. (2000). An Investigation of Teachers' Beliefs of Students' Algebra Development. *Cognition and Instruction, 18(2),* 209-237.

Newton, P. (1997). Examining standards over time. *Research Papers in Education,* 12(3), 227-248.

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 item 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.

Pollitt, A., Hutchinson, C., Entwhistle, N., & de Luca, C. (1985). *What makes examination items difficult?* Edinburgh: Scottish Academic Press.

Ofqual (2010). GCSE, GCE, principal learning and project code of practice. Available online at: http://www.ofqual.gov.uk/for-awarding-organisations/96-articles/247-codes-of-practice-2010 Accessed 26/07/10

Stewart, W. (2010). Grade boundary changes could leave results in doubt. In The Times Educational Supplement 21/05/10. Available online at: http://www.tes.co.uk/article.aspx?storycode=6044572 Accessed 23/07/10.

Thorndike, R. L. (1980). *Item and Score conversion by pooled judgements*. Paper presented at the Educational Testing Service Conference on Test Equating. Princeton, NJ.

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.

Wright, B. D. and Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Pres.