AWARDING OBJECTIVE TEST PAPERS: IS THERE A CORRECT ANSWER?

Lesley Meyer

CONTENTS

1. SUMMARY

In AQA, Objective Tests (OTs) have, in GCSEs generally, historically been awarded via a process known as 'back calculation' from other components (see §2.1). More recently, other methods such as Angoff (GCE) and KS3 predicted outcomes/IRT (GCSE Science) have also been employed. However, back calculation and the use of predicted outcomes do not involve professional judgement. This is in contrast to the Angoff approach, but this latter method has had mixed success as used in the current AQA GCEs. In 2004 and 2005, trials of awarding OTs using the Bookmark method were carried out in the AQA Manchester Research Department (Fowles, 2004 & 2005), but these did not give promising results and the approach was not pursued.

One OT unit (comprising one component – entirely OT) in the new Diploma qualifications is being awarded in January 2009, prompting the need for discussion on the most appropriate way to award this unit specifically, but also consequently for a review of the current methods available of awarding OTs generally to fully inform that decision. In the wider view, the awarding of OTs is gathering an increasingly high profile within Ofqual. This autumn (2008) the Standards and Technical Advisory Group (STAG) awarding subgroup is intending to draft a procedural paragraph for the JCQ Setting Standards papers[1] on the awarding of OTs, and Edexcel and OCR will be awarding new OT components in the Diploma in the forthcoming series[2], thus the awarding of this type of test is currently a hot topic both internally and externally.

This paper reviews the most recent literature on the principal current methods of awarding which can readily be applied to OTs (and which incorporate professional judgement into the process). The methods are compared, and the apparent pros and cons of each approach summarised, to provide an overall evaluation of the current scenario. The paper is also intended to serve as a basis for discussion as to how the award of the new AQA Diploma OT unit in January 2009, and those of other new OT units in the future, should be conducted.

2. BACKGROUND: EXISTING AQA APPROACHES TO AWARDING OBJECTIVE TEST COMPONENTS

The awarding of objective test, i.e. multiple choice, papers poses an interesting challenge. Since there is very little evidence on a completed script for an awarding committee to scrutinise which will provide any substantive evidence of candidate ability, other than the number of questions the candidate completed correctly, a judgemental scrutiny process similar to that used to award a standard written paper is inappropriate. Within AQA, over time, various different approaches have therefore been introduced to award OT papers.

2.1 Linear GCSE: Back calculation

For each judgemental grade of an examination, a weighted average of the cumulative percentage of candidates achieving that grade on the components in the examination *other than* the OT is calculated[3]. The boundary mark for the OT is then set at the mark corresponding to the cumulative percentage on the mark distribution for the OT which is closest to the calculated percentage.

[1] Setting Standards – Technical Issues and Setting Standards – The Awarding Process, which are updated annually by STAG.
[2] probably via "Angoff-type" and "equi-percentile to a written element-type" approaches, respectively, in the absence of a preferable option (personal communications from Jeremy Pritchard and Elizabeth Grey, 12/09/08). OCR's new GCSE examinations also involve OTs, which OCR intend to award via pre-testing.
[3] i.e. a proxy percentile method calculation based on all components excluding the OT.

Setting aside for now the lack of professional judgement in this approach, it could be argued as justifiable in the sense that the OT component outcomes are at least based on those of the other components, and the balance of outcomes across between the OT and the other components is therefore maintained. However, the obvious counter-argument is that candidates may perform very differently on an OT from a written paper or a coursework component, thus why should the grade boundaries for an OT be determined by the results on the other components?

2.2 Two-component GCE units: The Angoff method

While back-calculation is applicable in most linear examinations, it is less so in the modular scenario as most units/modules comprise a maximum of two components, the vast majority being single component only. In 2001, OTs were included in some of the (then new) AQA GCE examinations and, while an equi-percentile approach to grade boundary setting could have been employed as these OTs were all accompanied by a second component forming a two-component unit, it was desirable to incorporate professional judgement into the awarding process. The Angoff method (Angoff, 1971) was therefore introduced, being widely known and having been observed to be effective in the setting of Key Skills pass boundary marks at that time. Various modifications of the Angoff method exist, which are covered in more detail in §3.1.3, but the basic approach is as follows.

The Angoff method involves a panel of judges (awarders), who are asked independently to review each item included in the OT and provide an estimate of the proportion of 'minimally competent', i.e. borderline, candidates who would answer the question correctly. The individual item estimates are then summed and averaged across awarders to produce an overall percentage of items required to be correctly answered, which can then be translated into the minimum score required to pass at that grade boundary. In most applications (although not generally in AQA, see below) the participants provide more than one rating for each item - two rounds of rating are standard and usually no more than three. Between each round, the awarders are provided with feedback on, for example, their individual and group estimates and actual item performance indices.

In the current AQA GCE examinations, predicted outcomes, adjusting for candidates' prior achievement at GCSE, are used to maintain unit and subject standards at each series. Additionally, the Angoff approach is employed in various ways to inform the awarding of the GCE OT units, dependent mainly on the nature of the unit in which it is applied. For example, in GCE Chemistry the single OT (CHM6/W) combines with either a coursework component or practical to form each optional unit overall (CH6C and CH6P, respectively). Since the coursework and practical marks are anticipated to be carried forward from series to series, the OT boundaries have to be fine-tuned with an eye to the fact that, once the component marks are aggregated[4], it is desirable to achieve comparable outcomes across the two units (in relation to the unit-level predictions), while also ensuring that the overall subject outcomes are maintained within expected limits[5]. Consequently, only a single Angoff process is used to suggest an initial starting point for the OT boundaries and, given the context, there is limited discussion of the Angoff results *per se* in the awarding meeting.

[4] via the addition method, i.e. the sum of the scaled grade boundary marks across the components.
[5] These expected limits are termed 'guidance limits' in AQA and vary according to entry. For example, for a subject with 500 or more candidates entering this year and last year, the outcomes this series for matched candidates would be expected to be within 2% of those predicted. This limit increases as the size of entry decreases.

In GCE Economics there are two units which involve OT components (ECN1 and ECN2), each comprising the OT plus a written paper. A single Angoff process is used to provide initial estimates for the OT boundaries. Statistically equivalent boundaries on the written component are calculated and the aggregated unit-level mark resulting from these initial indications is then considered against the preferred mark suggested by the unit-level prediction in order to finalise the statistically recommended boundaries prior to the award. In the awarding meeting the written paper boundaries are considered first and the suggested OT boundaries then reviewed and finalised in the light of these written paper decisions with a view to maintaining the unit- and, ultimately, subject-level outcomes. GCE General Studies A also has various units comprising OT and written paper components and follows a very similar process to that used in Economics, although specific discussion of pre-test and facility information is included when finalising the General Studies A OT boundaries.

GCE Physics A (unit PAO4) also comprises an OT and a written paper, and is the only GCE unit in AQA for which a two-step Angoff process has been retained. Further, in the award the OT component is taken *first*: the initial Angoff results are discussed alongside additional facility information, and the Angoff process repeated to establish final OT marks. The written paper boundaries are then recommended considering the desired marks (according to the unit- and subject-level predictions) for the unit overall.

2.3 Single component GCSE Science modules and Functional Skills: Common candidates and common centres' data and/or KS3 predicted outcomes

While the AQA GCE units involving OTs all comprise two components, the OT modules in legacy and new specification GCSE Science were, and are currently, 100% OT. With twelve, tiered modules in legacy GCSE Science the Angoff approach was considered too unwieldy to be practically useful[6]. In the new specification the situation is little different, as there are six, tiered modules, all offered in written form and on-screen[7]. A statistical approach to awarding these legacy and new modules was therefore agreed with QCA (now Ofqual).

In legacy GCSE Science, common centres' mark distributions[8] were used alongside those for all candidates to suggest equivalent boundary marks in each series. Where there were differences between the marks suggested from the all candidate and common centres' data, common candidate data[9] were investigated to clarify which of the two marks would most closely maintain the balance of outcomes between modules. This was a time-consuming and involved process, but nonetheless logical, reproducible and understood by the examining team. The final boundaries resulting from the statistical analyses were reviewed ultimately by the Chair, and if necessary amended, before being put forward as the final recommendations.

By the time the new GCSE Science specification OTs were introduced, the use of predicted outcomes based on candidates' prior Key Stage 3 scores had become prevalent in modular and Applied GCSE specifications. These data are generally considered more reliable than common centres' and common candidate data – being rooted at candidate, rather than centre, level - and

[6] To complete the Angoff process, each awarder would have been required to supply the pass probabilities for thirty-six items, at two judgemental boundaries, at two tiers and for each of the twelve module tests, i.e. to provide 1,728 individual judgements!

[7] Although hitherto the written and on-screen versions have been awarded the same grade boundaries, there is no guarantee that this will continue.

[8] i.e. mark distributions based on candidates only from centres which have made entries this year and last year. These common centres would normally be considered the most stable in terms of their anticipated outcomes from year to year.

[9] i.e. mark distributions based on candidates entering more than one module.

have been used from the outset to guide the OT grade boundary setting in the new GCSE Science specification. A further refinement and extension to the supporting statistical analyses for this award is the use of IRT methods specifically to align standards at grade C across the two tiers. Common centres' and common candidate data are also used, where necessary, as further supportive information. Thus what the award lacks in the form of professional judgement is made up for in the depth of technical support involved which, although still extremely time-consuming, is more robust, reliable and reproducible than a judgemental method potentially based on individuals' estimates of over one thousand individual item probabilities.

The most recent OTs to be introduced and awarded are those in the pilot Functional Skills English and ICT specifications, in the February and June series in 2008. In English, the Reading unit is an OT, available at two tiers (Level 1 and Level 1/2), and in ICT there is one, non-tiered OT unit covering both Levels. A statistical approach has again been the main guide to the standard setting on these units, combining the use of predicted unit outcomes based on candidates' KS3 scores and an equi-percentile approach based on common candidates (candidates common to the OT and the written unit (unit 2), in each specification). The primary reason for the statistical approach was the need to assist the awarders in setting the (new) Level 1 and 2 standards, by informing them of the statistically recommended marks which would align the standards approximately to GCSE grades G and C on all the Functional Skills units, including the OTs (GCSE grades G and C being roughly equivalent to Levels 1 and 2, respectively).

With the introduction of the Diploma in 2009, an entirely new set of standards will need to be set, for the units at least, at Levels 1, 2 and 3. With no equivalent prior qualification on which to base standards the task is daunting, regardless of the form of the component being awarded. For OT components, however, the task is particularly challenging. One AQA OT Diploma unit (100% OT) is available in the first series, January 2009, prompting the need for agreement on the most appropriate way to establish grade boundaries for this unit, as well as others in the future. The recent literature on setting grade boundaries for OT components has therefore been reviewed to provide an informed basis for these discussions.

3. THE METHODS AVAILABLE

The main standard-setting approaches that are particularly well suited to awarding OTs fall into three main categories: absolute methods, item-mapping methods and compromise methods. Details of the different approaches and methods are summarised below. All are used in the USA for standard-setting, primarily in educational contexts (excepting the Nedelsky method, see §3.1.1), some more extensively than others. Consequently, descriptions of the methods tend to use terms such as 'raters', 'judges' and 'participants' rather than 'awarders' in relation to the teams of people involved, and refer to establishing 'cut scores' or 'cut points' rather than 'grade boundaries'. These alternative terms have been deliberately retained for the purposes of this report, if only to make clear throughout that these methods are not generally as prevalent in the UK. Only in the final discussion, when the potential implications for AQA's awards are discussed, is the more usual AQA awarding terminology reinstated.

3.1 Absolute methods

These methods are so termed because the subject experts are required to rate every item in the test without regard to normative information and performance data[10] (at least in the methods' original forms). Apart from the Jaegar method (§3.1.4), all require the evaluation of the difficulty

[10] i.e. mark distributions and item-level data, for example.

of items (or set of items in the Ebel approach, §3.1.2) with respect to the performance of a hypothetical group of minimally competent candidates.

3.1.1 The Nedelsky method

The Nedelsky method, developed in 1954 and cited by Cizek and Bunch (2007), Cizek (2001) and Mehrens (1995), amongst others, involves assigning values to multiple-choice test items based on the likelihood of examinees being able to rule out incorrect options. The standard-setting participants inspect the test items and identify, for each item in the test, any options that a hypothetically minimally competent candidate would rule out as incorrect. The reciprocal of the remaining number of options becomes each item's 'Nedelsky rating', i.e. the probability that the minimally competent candidate would answer the item correctly. For example, on a five-option item for which a participant considers candidates would rule out two of the options as incorrect, that participant's Nedelsky rating for that item would be 1/(3 remaining options)=0.33. The mean Nedelsky value across the raters is calculated for each item and the sum of these means[11] is used as a passing score.

Hypothetical data illustrating this method for a fifteen item test with six participants is shown in Table 1 (adapted from Cizek & Bunch, 2007, page 72). Where the passing score is not a whole number the passing score is rounded up, thus in this example only candidates who attained 7 raw marks (or higher) would pass at that grade.

Table 1: Nedelsky's method

	Rater number and Nedelsky values						
Item	**1**	**2**	**3**	**4**	**5**	**6**	**Item means**
1	0.33	0.50	0.50	0.33	0.33	0.33	0.39
2	0.50	1.00	0.50	0.25	1.00	1.00	0.71
3	0.25	0.33	0.25	0.25	0.25	0.33	0.28
4	1.00	1.00	0.50	1.00	0.50	1.00	0.83
5	0.33	0.33	0.33	0.33	0.25	0.33	0.32
6	0.25	0.33	0.25	0.25	0.25	0.33	0.28
7	0.25	0.20	0.25	0.33	0.20	0.20	0.24
8	1.00	0.33	1.00	0.50	1.00	0.50	0.72
9	0.20	0.33	0.25	0.20	0.33	0.25	0.26
10	0.50	1.00	1.00	0.50	0.50	1.00	0.75
11	0.50	0.50	0.50	1.00	0.50	0.50	0.58
12	0.50	0.33	0.33	0.50	0.33	0.33	0.39
13	0.20	0.20	0.20	0.20	0.20	0.20	0.20
14	0.25	0.20	0.33	0.25	0.33	0.25	0.27
15	1.00	0.50	0.50	0.50	1.00	0.50	0.67
Sum of item means (rounded up to produce the final cut score):							**6.87**

The Nedelsky method is popular in medical contexts, presumably (Cizek, 2001, suggests) because the minimally competent practitioner should be able to reject those options that would cause harm to patients(!). However, serious limitations of this method have been described in the literature. It can only be applied to the multiple choice format (although that in itself does

[11] Nedelsky also suggested an alternative to the basic procedure, which adjusts the simple sum of the ratings to take into account the kinds of item choices within the test: when the test contains a large proportion of answer choices which the borderline student will recognise as incorrect a greater adjustment is made than for a test with comparatively fewer clearly incorrect answer choices, thus fewer borderline candidates would be allowed to pass in the former scenario than in the latter. However, this modification is rarely, if ever, applied in practice.

not concern us for the purposes of this paper). More importantly, ostensibly it cannot be used in situations where more than one cut score is required on the same test and there do not appear to have been any suggested modifications allowing for such a scenario, although it should not be impossible to accommodate such changes. (Obviously, to set two cut scores on the same test raters would have to provide two ratings for each item.) Another weakness of the Nedelsky method is that there are a limited number of probabilities raters can assign to an item. For example, for a five-option item, the only Nedelsky values that could result from a participant's ratings would be 0.20, 0.25, 0.33, 0.50 and 1.00. Further, there are not equal intervals between these probabilities. Since, in practice, raters tend not to assign probabilities of 1.00 (i.e. to judge that a borderline candidate could rule out all incorrect responses), this tends to create a downward bias in item ratings (as 0.50 would be assigned to the item rather than 1.00, for example). The overall cut score is therefore lower than the participants may have intended to recommend and consequently the method tends to produce somewhat more lenient cut scores than other methods.

3.1.2 The Ebel method

Another item-based approach is that proposed by Ebel (1972, pages 492-496), in which the standard-setting participants have to provide two judgements for each item: one being an estimate of the difficulty of each item, the other regarding the relevance of the item. Instead of participants expressing item judgements in the form of probabilities, they are instead required to classify items into, most commonly, three difficulty levels (easy, medium, hard) and four relevance levels (essential, important, acceptable, questionable), resulting in twelve difficulty-by-relevance combinations overall. At this stage these judgements are not made with respect to a hypothetical candidate, rather with respect to the purpose of the test and the overall candidate population. Thus the participants are simply required independently to judge whether they believe the items are easy, medium or hard for the examinees as a group, and whether the individual items are essential, etc. with respect to the grade boundary being set. The participants are then asked to make a further judgement, this time considering how minimally competent candidates will perform on the test: for each difficulty-by-relevance combination they are asked to estimate the percentage of items that a minimally-competent candidate should answer correctly. (This latter judgement is often carried out via group discussion to arrive at a consensus opinion, rather than each participant providing independent estimates in isolation.) To obtain a cut score, the number of items judged to be in each difficulty-by-relevance combination is multiplied by the percentage of those items that the participants have estimated should be answered correctly. These products are summed and divided by the total number of judgements involved to give an estimated percentage correct and thereby a cut score.

The data in Table 2 (adapted from Cizek & Bunch, 2007, page 77) illustrate the application of the Ebel method. Hypothetical classifications are shown for 100 items made by a panel of five participants, yielding a total of 500 judgements (each item is categorised by each judge). Here, 94 of the relevance judgements classified items as 'Essential', 259 as 'Important', 124 as 'Acceptable' and 22 as 'Questionable'. In terms of item difficulty, 228 items were judged to be 'Easy' (94+106+24+4), 213 'Medium' and 59 'Hard'. The participants agreed that a minimally-competent candidate should answer all the 'Essential' items correctly whatever the difficulty category, whereas in the other relevance categories the candidate would be expected to answer correctly a higher percentage of easy items than hard ones. The judgements yield a recommended 75.26% passing percentage correct, i.e. approximately 75 or 76 of the 100 items must be correct in order for a candidate to pass (at this grade).

This method could be adapted when item performance data are available such that, rather than arbitrarily categorising items as Easy, Medium or Hard, participants would decide on the borderlines of these three categories according to the item performance data. For example, items with facility indices of 0.00 to 0.49 could be classified as Hard, 0.50 to 0.79 as Medium and 0.80 or higher as Easy. Another amendment could be to ask each judge independently to estimate the percentage correct for each relevance-by-difficulty combination and then calculate, say, the mean of these judgements to determine the final judged percentage correct, rather than determining the value by consensus.

Table 2: The Ebel method

Relevance category	Difficulty category	No. items judged to be in category (A)	Judged percentage correct (B)	Product (A x B)
Essential	Easy	94	100%	9, 400
	Medium	0	100%	0
	Hard	0	100%	0
	Subtotal	94		
Important	Easy	106	90%	9, 540
	Medium	153	70%	10, 710
	Hard	0	50%	0
	Subtotal	259		
Acceptable	Easy	24	80%	1, 920
	Medium	49	60%	2, 940
	Hard	52	40%	2, 080
	Subtotal	125		
Questionable	Easy	4	70%	280
	Medium	11	50%	550
	Hard	7	30%	210
	Subtotal	22		
TOTALS		**500**		**37, 630**
			Percentage passing: 37, 630/500=	**75.26%**

The method is readily used for dichotomously scored items, but can also be used for polytomous items and is not restricted to multiple choice formats. However, although it is adaptable and easily implemented, this method has received criticism. For participants, keeping the two dimensions of difficulty and relevance distinct may be difficult, particularly as in some situations these may be highly correlated. Asking participants to categorise items into difficulty categories may not be necessary if item information is available, but the use of such 'real time' data could cause difficulties if the participants' views of the difficulty levels differ markedly from the known values. Another concern is that the method, in effect, reveals inadequacies in the test-construction process (why, for example, should any items of questionable relevance be included in any examination?).

3.1.3 The Angoff method

Originally proposed by Angoff (1971) and described in detail in Cizek and Bunch (2007), amongst others, this method is the most thoroughly researched of the item-based standard setting procedures currently available and is the most widely used, having been adapted into many different variations. Angoff's original proposal suggested that:

> *A systematic procedure for deciding on the minimum raw scores for passing and honors might be developed as follows: keeping the hypothetical and "minimally acceptable person" in mind, one could go through the test item by item and decide whether such a person could answer correctly each item under consideration. If a score of one is given for each item answered correctly by the hypothetical person and a score of zero is given for each item answered incorrectly by that person, the sum of the item scores will equal the raw score earned by the "minimally acceptable person". (Angoff, 1971, 514-515).*

However, in practice a footnoted variation to the proposal has dominated applications of this method:

> *A slight variation of this procedure is to ask each judge to state the probability that the "minimally acceptable person" would answer each item correctly. In effect, judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score. (Angoff, 1971, 515).*

a) The 'Modified Angoff' method

In most applications Angoff's footnoted variation of the procedure is modified to facilitate less variable estimations of the final boundary mark. In these 'modified Angoff' approaches, two rounds of rating are standard and usually there are no more than three. In between rounds the participants are provided with normative data – typically feedback on, for example, their individual and group estimates, allowing the opportunity for participants to see how their ratings compare with those of others, and actual item performance indices. The revised item estimates following the provision of this information do not have to change, but in practice they tend to converge as the item probability estimates are more accurate, the amount of between-participant variation is reduced, and/or the participants become aware of the consequences of their ratings in the context of the resultant cumulative percentage passing at that judgemental grade. Once the judges have provided their final ratings and the sum of these across items has been calculated for each judge, the mean average across the judges is computed to establish the final grade boundary. (In AQA, the highest and lowest of the awarders' totals are omitted prior to calculation of the mean, which is then rounded down to establish the final boundary mark – a further variation of the Angoff approach inherited from QCA's awarding of Key Stage 1 tests in 2000.) An example of the application of such a modified Angoff method in an AQA GCE OT component, with fifteen items and involving five awarders giving two iterations of judgements on each item, is given in Table 3[12]. Note that, in the AQA adaptation, the awarders are asked to consider a group of 100 candidates who are just inside the grade boundary in question and decide how many of them would answer the item correctly, hence the estimate of 35 from awarder 1 in round 1 in essence implies that awarder considered 35% of candidates would provide the correct answer to that item.

(It should also be noted that in the literature the calculations on the awarders' estimates are normally shown as follows: after each round the average of each awarder's estimates is calculated, along with the overall average across all items and awarders. This gives a recommended passing percentage which is then converted to a cut score. For example, using the data in Table 3, the averages of the awarders' estimates following round 2 are 40.3, 41.7,

[12] with many thanks to the Subject Officer for sharing these data.

45.0, 44.1 and 43.7, for awarders 1 to 5, respectively, and thus the overall mean estimated percentage correct would be 43.0. Given that there are 15 items on the test, this implies a proxy grade boundary of 6.45 (i.e. between 6 and 7 items must be correct to pass). Ignoring the highest and lowest awarder average (as in AQA's inherited approach) would yield an overall average of 43.15% correct, i.e. a proxy grade boundary of 6.47, as shown in Table 3.)

Table 3: The Angoff method (AQA's variation)

Item and Round		Awarder 1	2	3	4	5	Average across awarders	
1	Round 1	35	35	50	48	45	42.60	
	Round 2	40	55	55	53	70	54.60	
2	Round 1	30	30	40	25	19	28.80	
	Round 2	40	35	45	33	32	37.00	
3	Round 1	55	30	50	40	46	44.20	
	Round 2	45	35	50	42	42	42.80	
4	Round 1	40	45	45	63	22	43.00	
	Round 2	45	50	50	63	50	51.60	
5	Round 1	45	40	50	36	23	38.80	
	Round 2	45	45	55	44	42	46.20	
6	Round 1	30	35	45	47	16	34.60	
	Round 2	40	40	60	52	58	50.00	
7	Round 1	35	30	35	28	42	34.00	
	Round 2	40	45	35	29	41	38.00	
8	Round 1	40	45	45	16	17	32.60	
	Round 2	45	50	40	25	46	41.20	
9	Round 1	40	35	35	45	14	33.80	
	Round 2	40	40	35	48	36	39.80	
10	Round 1	45	40	40	48	45	43.60	
	Round 2	45	45	45	49	38	44.40	
11	Round 1	30	30	40	37	49	37.20	
	Round 2	25	40	40	37	38	34.20	
12	Round 1	40	35	40	51	45	42.20	
	Round 2	30	40	45	51	36	42.40	
13	Round 1	40	40	35	39	29	36.60	
	Round 2	30	30	35	35	34	32.80	
14	Round 1	50	40	40	41	34	41.00	
	Round 2	45	40	40	43	38	41.20	
15	Round 1	35	35	45	54	34	40.60	
	Round 2	40	45	45	57	54	48.20	
Total/100	Round 1	5.90	5.45	6.35	6.18	4.80	Mean* Round 1:	5.84
	Round 2	6.05	6.25	6.75	6.62	6.55	**Mean* Round 2:**	**6.47**
							⇒ grade boundary = 6	

*across awarders, omitting the highest and lowest awarder totals

b) The 'Yes/No' method

Although most applications use Angoff's variation on his original proposal (i.e. requiring each judge to state the probability that the 'minimally acceptable person' would answer each item correctly), there is evidence that judges find this a difficult task (Impara & Plake, 1997) and the issue has cast doubt on the validity of item-based procedures generally (see later discussion in §3.1.3(d)). An alternative Angoff approach, which follows more closely Angoff's original proposal, is therefore also used. Termed the 'Yes/No method', it reduces the cognitive task facing the judges by requiring the judges to think of a borderline student and judge whether that student would be able to answer each item correctly (i.e. reducing the probability estimation to a dichotomous outcome, while also reducing the conceptual burden on the judges by asking them to focus on a typical borderline candidate rather than a hypothetical group of borderline candidates). Typically two rounds of ratings are still involved with feedback in between, as per

the standard Angoff approach. This approach is appealing because of its simplicity (Cizek, Bunch & Koons, 2004) and, in their 1997 account of two studies, Impara and Plake considered the method to have "substantial promise". The panellists found the method clearer and easier to use than the more traditional Angoff approach, finding it simpler to think of an actual student as opposed to a group of hypothetical students. Further, while the Yes/No method gave similar overall results to the more conventional Angoff approach, the variance of the ratings was smaller with the Yes/No method and the participants' scores more stable across rounds one and two. Overall Impara and Plake conclude that the performance standard derived form the Yes/No method may therefore be more valid than that derived from the traditional Angoff method. Chinn and Hertz (2002) also reported that participants found decisions easier to make via the Yes/No method but, in contrast to Impara and Plake, found that there was greater variance in the ratings than with the standard Angoff approach. However, the Chinn and Hertz study has various flaws: most particularly they did not ask judges to identify a specific candidate when making their ratings, but also the allocation of judges to judgemental groups was non-random. A further potential weakness of the Yes/No method, flagged by Cizek and Bunch (2007), is the potential for either positive or negative bias in the item ratings, depending on the clustering of item difficulty values in the test. The potential for bias arises because the method is based on an implicit judgement of whether the probability of a correct response at the cut score is greater than 0.50. To take an extreme example, suppose that a test comprised identical items that all had a probability of correct response at the cut score of 0.70. An accurate rater would assign ratings of '1' ('Yes') to each item and the resulting performance standard would be the maximum mark for the test, which is unlikely to be what the rater intended, nor a realistic expectation based on the difficulty of the test.

c) The 'Extended Angoff' method

Another well-known adaptation of the Angoff method exists for constructed-response items (i.e. items requiring either a short, or extended, written response) and is termed the 'Extended Angoff' method (Hambleton & Plake, 1995). Instead of providing conventional probability estimates of borderline examinee performance for each item in the test, panellists are instead asked to estimate the typical score that a borderline examinee would be expected to obtain on each question. The averages of the panellists' performance estimates for each question on the test are then calculated and aggregated to an overall mean across participants to yield the expected performance standards on the test overall and thus the potential cut score. Although less directly relevant to the purposes of the current paper it is worth mentioning, as the obvious benefit of this extension is that it enables the Angoff approach (in various modifications) to be used in a test comprising mixed item formats, should that scenario present itself. A detailed example of the application of the Extended Angoff method is given in Cizek and Bunch (2007), pages 87 and 88.

d) General discussion on the Angoff approaches

To date, the literature suggests that the Angoff method (or a modification of it) is still a preferred approach, which is not surprising as the research indicates it provides results which are easy-to-obtain and acceptable in many situations (Cizek, 1996). To summarise but a few studies, Cizek and Bunch (2007, page 82) cite a study of the Angoff, Ebel and Nedelsky methods by Colton and Hecht (presented at the National Council on Measurement in Education in 1981) which considered the Angoff approach and consensus technique to be superior. Cross, Impara, Frary and Jaegar (1984) compared the Angoff, Nedelsky and Jaegar approaches and concluded that the Angoff approach produced the most defensible standards, noting also that the participants expressed greater confidence in their Angoff estimates and in the standards resulting from their judgements than those of the other two methods. In his review of twenty-three standard-setting

methods, including those of Nedelsky, Ebel, Beuk (see §3.3.2) and Hofstee (§3.3.1), Berk (1986) also concluded that the Angoff method appeared to offer the best balance between technical adequacy and practicability. Mills and Meilican (1988) commented that the method appeared to be the most widely used, having the benefit that it is not difficult to explain, as well as the fact that the data collection and analysis are comparatively straightforward and simpler than for other similar approaches. Further still, in Mehrens' 1995 review of the literature, Angoff's method is again suggested to be the preferred and recommended model (in comparison to those of Nedelsky and Ebel), based on the reasonableness of the standard set, ease of use and the fact that the standard is more reliable, the inter-rater and inter-judge consistency being higher. Another benefit of the Angoff approach, which is particularly relevant to AQA, is that it can be readily modified to allow the setting of more than one cut score on a test (Cizek & Bunch, 2007).

Making judgements about the performance of borderline students is central to the Angoff method, as well as those of Nedelsky and Ebel. However, whether or not participants can actually conceptualise such students and estimate their performance on test items has become an issue of controversy. Cizek and Bunch (2007, page 95) cite a 1993 report from the National Academy of Education (NAE) on the implementation of a modified Angoff approach in setting standards for the National Assessment of Educational Progress (NAEP), which provided some evidence related to the inability of standard-setting participants to form and maintain the kinds of conceptualisations required to implement item-based procedures. The report suggested that abstractions such as minimally-competent or borderline candidates may be impossible for participants to acquire, and to adhere to once acquired. It concluded that the Angoff procedure was "fundamentally flawed" and that the use of the Angoff method, or any item-judgement method to set achievement levels should be discontinued[13]. In the introduction to their 1997 study of the 'Yes/No' Angoff approach, Impara and Plake (page 354) cite various earlier reports which also question the ability of judges to estimate item difficulty accurately and their own research the following year concurred (Impara & Plake, 1998). They too consider that the confidence which can be placed on the accuracy of an estimated proportion correct may be low, even when the judges have a high degree of familiarity with both the examinees and the test. They suggest that judges may have difficulty in conceptualising hypothetical borderline candidates and also that estimating the proportion correct may be very difficult, even for a clearly defined group of examinees. In this context, the 'Yes/No' method, which addresses both issues, is therefore a very sensible development. Later research by Clauser, Swanson and Harik (2002) examined the impact of training and feedback on various sources of error in the estimation of cut scores in an Angoff-style standard setting procedure. Their results indicated that after training there was little improvement in the ability of judges to rank order items by difficulty, but there was a substantial improvement in inter-judge consistency in centring ratings.

Nevertheless, to date little empirical attention has been paid to the various accusations against the Angoff method. Particularly, the criticisms made in the 1993 NAE report have been strongly refuted by leading psychometricians with expertise in standard-setting, who have argued robustly that the report is one-sided, incomplete, based largely on dated and second-hand evidence and "lacking in scientific credibility" (Hambleton *et al*., 2000, page 8). Further, Clauser *et al*. (2002) point out that while accurately rating item difficulty is a challenging task, this in itself provides no evidence that the Angoff procedure is flawed. Rather the difficulty of the activity highlights the importance of incorporating examinee performance information and feedback in the standard-setting procedure. Throughout the late 1990s and beyond the Angoff method has

[13] It is worth pointing out, however, that the abstract concept of a borderline candidate is accepted for awarding purposes in the UK, e.g. within the standard AQA scrutiny and tick chart procedures for awarding written papers.

remained the most recommended, preferred and prominent cutscore method available (Mehrens, 1995; Hurtz & Hertz, 1999; Zieky, 2001). The widespread use of item-judgement methods and the Angoff method in particular therefore seems likely to continue in the future.

Indeed, very recently, Béguin, Kremers and Alberts (2008) trialled an innovative new variant on the Angoff method, in the context of transposing standards from an old examination to a new one. A set of items was rated, containing items from both the old and the new examinations, the old items (*i=1, 2,…, k*) functioning as an anchor for the standard on the old examination. The raters were asked to estimate the probability of a correct answer on each item in the total population of candidates[14]. Then the severity (or leniency), s_r, of each rater, *r,* on the *k* anchor items is defined by $s_r = \sum_{i=1}^{k} (\hat{p}_{i,r} - p_{i,r}^{obs}) / k$, where $p_{i,r}^{obs}$ is the observed probability correct for the anchor item *i* in the population of candidates and $\hat{p}_{i,r}$ is the estimated probability of a correct answer on item *i* by rater *r*. Subsequently, for each new item, the probability corrected for severity, $p_{i,r}^{c}$, is determined by $p_{i,r}^{c} = \hat{p}_{i,r} - s_r$. Taking the average of these corrected probabilities for the new items over the items and over the raters the expected difficulty of the new examination (and thereby the new cut point) is determined. Although the procedure was easy to apply, Béguin *et al*. found it not to be effective in their recent study. As an indicator of effectiveness, they calculated the reduction in variance between the corrected and uncorrected ratings but, after correction, rather than decreasing the standard deviation between the raters increased (from 4.67 to 5.13 marks), implying that the raters' levels of severity (or leniency) did not remain consistent across the items. Thus, although in theory it may have potential, the method would need further investigation and development before being considered for operational use.

3.1.4 Jaeger's method

A further item-based approach similar to Angoff's was developed by Jaegar (1982). This procedure avoids the potentially troublesome definition of what constitutes 'minimally competent' by requiring participants to answer the following question for each item in the examination: "Should *every* candidate…be able to answer the test item correctly?" The number of items marked as "Yes", is summed for each judge to give that judge's recommended cut score (at that point in time). Similar to the various Angoff approaches, Jaegar's procedure requires (usually three) iterations of data collection, with participants being provided with an opportunity to reconsider their judgements in each iteration. Between each stage of judgements they receive data on the standards currently implied by their own Yes/No ratings and those of the group as a whole, as well as item difficulty data based on actual candidate performance. The minimum median cut score across the judges after session three determines the final cut score.

Similar to the Nedelsky method, Jaegar's approach has been criticised for being time-consuming and limiting the participants' probability choices to 0 or 1 (Berk, 1986). It may also produce less reliable standards than other item-based approaches (Cross *et al*., 1984). However, the relative youth of the Jaegar method means that to date it has received less scrutiny than those of Angoff, Ebel and Nedelsky.

[14] As opposed to estimating the probability of a correct answer in the population of 'just sufficiently proficient candidates' – the test construction teams in this study were more familiar with the concept of the full candidate population and therefore preferred to make estimates for this latter population.

3.2 Bookmarking

While the previous four methods all require judgements on individual items in a test, the Bookmark method (Mitzel, Lewis, Patz & Green, 2001) is designed to yield cut scores on the basis of participants' reviews of collections of test items. The method and its several variants, grew out of a series of 'item-mapping' strategies developed in the 1990s, which essentially 'map' items on to a proficiency distribution where cut scores are set. The main Bookmark approach and a variant of it are discussed below.

3.2.1 The Bookmark method

The Bookmark method is so-called because participants express their judgements by entering markers in a specially designed 'ordered item booklet' consisting of the set of items placed in order of difficulty, from the easiest item to the hardest. Detailed descriptions of the Bookmark approach can be found in various sources including Mitzel, Lewis, Patz and Green (2001), Cizek, Bunch and Koons (2004), and Cizek and Bunch (2007), but put simply the approach is as follows: the candidates' performance on each item determines an ordering of the items in terms of item difficulty (usually calculated according to item response theory (IRT), but alternatively via classical test theory (CTT), see below). The items are presented in ascending order of difficulty in an 'item booklet' (one item per page, see example in Figure 1) to a team of judges whose task is to identify the point in the booklet between the two items which they judge to mark the transition from candidates being proficient to not proficient. Proficiency is normally taken as 67% probability of success, although whether this is the most appropriate response probability is the cause of much debate (again, see below). Usually, the last item regarded as likely to be correctly answered is the item against which the judges are asked to set their bookmark. As with other standard-setting procedures, most applications incorporate two or more rounds of judgements, so that the judges can be given feedback, time for discussion and the opportunity to revise their decisions. The information about the bookmarked items is collected from each judge and used to estimate the average ability of the borderline candidates. The boundary mark is then derived using either the mean or median of all the judgements[15]. (The calculations are initially carried out based on the ability estimates as represented in the IRT model, but are ultimately converted to raw scores for the purposes of the awarders.) Hypothetical data illustrating how the awarders' judgements might appear following the first round of the Bookmark procedure are shown in Appendix A.

(It is important to note that the position at which an individual participant places their bookmark does not, in itself, establish their suggested cut score. For example, if a participant places the bookmark on page 20 of a 50-page ordered item booklet (OIB) to distinguish between Fail and Pass grades, this does not correspond to a pass mark of 20. This mark indicates that, in the participant's judgement, candidates would be expected to answer the questions correctly (with whatever probability of success is being used) on all the items through to page 20 of the OIB. To obtain the cut score indicated by that participant, it is necessary to establish the ability level associated with a response probability of 0.67 (or whatever probability is being used) that corresponds with the page in the OIB on which the bookmark was placed (here page 20). This gives that participant's cut score in 'ability scale' units. This process is repeated for each participant. The overall recommended cut score (in ability units) is derived by taking (usually) the mean of the recommended cut scores across all participants (see Appendix A). This is then converted to a final overall cut score on the raw mark scale via the IRT model.)

[15] Whether to use the mean or median has been suggested by Karantonis and Sireci (2006) as an area requiring further research (and is also relevant to other standard-setting methods). The mean is traditionally used, but using the median would minimize the influence of outlying judges – on the other hand it may be better to let such differences of opinion remain influential.

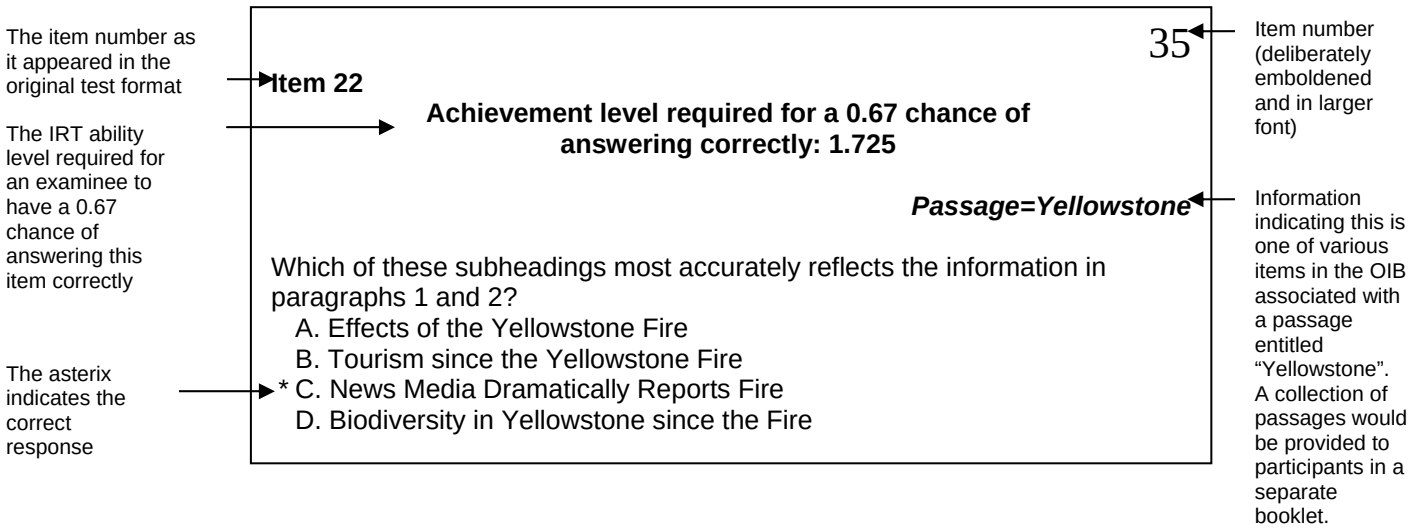

Figure 1: The Bookmark method - sample page from an ordered item booklet
(Source: Cizek, Bunch & Koons, 2004)

This approach has become quite popular for various reasons: it can be used for complex, mixed format assessments, for example a test comprising multiple choice and/or constructed response items, and multiple cut scores can be set for a single test. Also the judgemental task is simplified from that of the Absolute methods and the cognitive load is reduced: the panellists are provided with known difficulty information rather than having to estimate it themselves, the items are pre-ordered to help the participants understand the relative difficulties of the items and fewer judgements have to be made. Although the item difficulties are usually calculated according to IRT, CTT has also been considered which has the benefit of simplicity, accessibility and no strong assumptions which might be invalidated (Buckendahl, Smith, Impara & Plake, 2002; Schagen & Bradshaw, 2003). Schagen and Bradshaw considered CTT facilities alongside item difficulties obtained from a simple Rasch model and pointed out that, although there may be a generally monotonic relationship between the two approaches (classical facilities versus Rasch difficulties), some items, particularly multi-mark items, may counteract this trend and therefore the two models may result in different item orderings. Also, classical facilities are affected by pupils 'not reaching' items in the test due to time constraints: firstly, they are reduced if questions that have been omitted are treated as incorrect (as some of the candidates who did not reach the item may in fact have answered it correctly); secondly, they are inflated because more of the candidates who do reach the item will be of higher ability. Schagen and Bradshaw suggest the best way of allowing for both these effects is to use 'weighted' facilities, based on dividing candidates into groups according to a measure of ability, but also point out that correcting for 'not reached' in IRT software is relatively straightforward, as the final items for candidates who did not reach the end of the test can be coded as 'not presented' and the item and person parameters are then estimated excluding these cases.

Aside from the method of calculating the facilities, there has been much recent debate over other technicalities inherent to the Bookmark procedure: particularly whether 0.67 should be used as the probability of success, or whether 0.50 (or another alternative response probability (RP)) is preferable, and also whether a one-parameter (Rasch) model, or two- or three-parameter model should be used. Both of these are central to the Bookmark procedure,

affecting the ordering of the items in the item-booklet and the determination of the final cut score[16].

The developers of the Bookmark method suggested using a three-parameter logistic model for selected-response (SR) items and a two-parameter partial credit model for constructed-response (CR) items (Mitzel *et al*., 2001). However, a one-parameter logistic (Rasch) model is also frequently used in practice for SR and CR items where the latter are scored dichotomously. Beretvas (2004) explored the impact of choosing different IRT models and different response probability (RP) values on the ordering of items in the item booklet using the Bookmark approach and her results demonstrated how the use of different models and response probability values results in different item ordering. The Rasch model results in a different ordering from two- and three-parameter models. Further, for these latter models using different RP values results in different item ordering. Only when the Rasch model is used with dichotomously scored items does the item ordering remain the same regardless of the RP value used. Beretvas did not investigate the impact the variations of item ordering could have on the determination of the final cut score but it would seem likely that, given different item orderings, judges' bookmark placements would differ and therefore potentially so would their recommended cut scores. On the other hand, given the likely variation in where judges set their bookmarks, with the final cut score being the average of the ability estimates at those bookmark points, it may be that the effect of slightly differing rank orders of the items would not have an great effect on the cut score. This area deserves further research.

A probability of 0.67 is typically employed for the Bookmark procedure and two reasons have been provided to support this. First, participants are asked to place a bookmark in such a way as to separate items that a borderline candidate has mastered from those that they have not mastered. Mitzel *et al*. (2001) argue that a success probability of 0.67 is consistent with 'mastery' (in that it is greater than 0.50) and it is a relatively easy value for participants to understand. A further justification for using a value of 0.67 is based on research by Huynh (in 1998 and 2000, cited by Karantonis & Sireci, 2006, page 8) suggesting that the RP which maximised the item information function of the test would produce the optimal decision rule and, when guessing is removed under the three-parameter logistic model, the information function is maximised when the probability of a correct response is 0.67 (and when the ability of the candidate is equal to the item difficulty). However, Wang (2003) advocates using a probability of success of 0.50 in the Rasch model, as the item information function is maximised with this RP (see also §3.2.2) and Karantonis and Sireci (page 7) cite other unpublished research by Kolstad from 1998 which also supports the use of 0.50 as the RP criterion. Issues related to the selection of the most appropriate response probability to use therefore remain and, aside from the technical intricacies, there are practical issues fundamental to these arguments. Can standard-setting participants use any particular response probability value more effectively than another? Can they understand and apply the concept of a response probability more consistently and accurately than they can generate probability estimates using, for example, an Angoff approach? The results of a study by the National Academies of Sciences (2005) and also work presented at the annual meeting of the National Council on Measurement in Education in April of that year by Williams and Schultz (cited by Karantonis & Sireci, page 8) both suggest that participants find it more difficult to implement an RP value of 0.50 than 0.67

[16] Discussion of the AQA Bookmarking trials (see §4.6) at the November 2003 Research Committee led to a recommendation that IRT methodology in AQA, and hence the analysis underlying the Bookmark approach, should use only the one-parameter model, rather than two- or three- parameter models, as in the one parameter model there is a unique (monotonic) equivalence between candidates' marks and ability which is not the case in the other models. Nevertheless, the ongoing debate surrounding model selection in the wider context is included here for completeness.

but more research is needed to clarify the most appropriate RP value – both from a technical and practical perspective.

An additional concern which is the subject of ongoing research is the suggestion from various sources that the Bookmark method may systematically underestimate cut scores. Green, Trimble and Lewis (2003) compared the Bookmark procedure to two other standard setting methods (Jaegar-Mills and the Contrasting Groups method, which are not directly applicable to multiple choice tests and are therefore not discussed in this paper) and found that the Bookmark approach tended to produce the lowest cut scores. Karantonis and Sireci (2006, page 8) cite a paper by Reckase, presented at the April 2005 meeting of the National Council on Measurement in Education in Montreal, Canada, in which, via a series of simulations, he showed that even under ideal (error-free) conditions, the results of the first round of Bookmark cut scores were significantly negatively biased. Furthermore, when he simulated the error in the panellists' judgements the magnitude of the bias increased. Corroborative conclusions stemming from comparisons specifically between Bookmark-based and Angoff-based approaches are discussed in §4.4.

3.2.2 Item mapping

Item mapping is a variant on the standard Bookmarking procedure which presents an overall picture of all the items and their estimated difficulties in the form of a histogram (item map), which serves to guide and simplify the judges' decision-making process. As detailed by Wang (2003) the approach uses a one-parameter Rasch model to order the items and a RP of 0.50. The approach utilises the fact that, in the Rasch model, the difference between a candidate's ability and an item's difficulty (estimated on the same scale by the model) determines the probability of a correct response to the item. When candidate ability equals item difficulty the probability of a correct answer to the item is 0.50. A candidate with an ability level lower than the item difficulty will have less than 0.50 probability of success and one with an ability level higher than the item difficulty will have greater than a 0.50 probability of success. By utilising this feature of the Rasch model, the item-mapping method enables judges to establish the pass mark at the point where the borderline candidate's ability level equals item difficulty. As, when using a RP of 0.50 in the Rasch model, candidate ability is equal to item difficulty, from the judges' perspective, no mathematical adjustment is needed to establish a minimal competency point following their judgements on the items. If a different RP was used for item mapping (e.g. 0.67) the difference between ability and item difficulty corresponding to the RP of 0.67 would not be zero, and mathematical adjustments would need to be made to identify the location of minimal competency once the judges had established their estimated cut point (see Appendix B for the theoretical basis of this discussion).

A linear transformation is applied to the item difficulties from the Rasch model to translate them onto an arbitrary scale, but one that is more meaningful to the judges than the logit scale (on which the estimates from the model are based). Simultaneously, this transformation serves to group some adjacent item difficulties to narrow the range and thus enable the histogram to fit onto a single page. For example, the item difficulties tabulated in Appendix C(i) were first multiplied by 4 and then 40 was added to the product. The resulting transformed difficulties were rounded to the nearest integer and then plotted (see Appendix C(ii)) with the height of the columns being the number of test items on the rounded difficulty value. The items associated with each difficulty score are specified in each column of the histogram.

Working across the columns, the judges are asked to consider each item in that column and independently determine whether a typical borderline candidate has at least a 0.50 chance of

answering the item correctly. Group discussion then ensues to establish consensus as to the column of items where the probability of a correct response from a borderline candidate is 0.50 for most of the items. The corresponding middle level of item difficulty under this column is then translated (via the Rasch model) into the associated cut score.

The item mapping approach still requires the judges to be familiar with the items, thus preparation of an ordered item booklet (sorted in Rasch difficulty order) is probably still worthwhile, since with items of similar difficulty printed alongside each other in the OIB judges would not have to search through pages of the booklet to locate the items appearing together in one column. However, it is not imperative to the method and therefore, particularly for short tests, referring to the OT paper in its original form may be adequate.

3.3 Compromise methods

In the 1980s, various approaches were developed intending to build on those used by the Absolute methods by attempting to strike a compromise between purely norm-referenced (relative) approaches and absolute methods. The two main approaches are summarised below (more details can be found in Mills & Meilican, 1988; Cizek, 1996; and Cizek & Bunch, 2007, amongst others).

3.3.1 Hofstee's method

Hofstee's method is conceptually quite simple and is also very time-efficient. The judges are asked to respond to four questions and to assume that the examinees are taking the test for the first time. Two of the questions focus on the acceptable level of knowledge that the examinees should possess; the other two focus on the tolerable failure rates for the examination.

Question 1: "What is the highest percent correct cut score that would be acceptable, even if every examinee attains that score?" (This value is a participant's estimate of the maximum level of knowledge that should be required of examinees, k_{max}.)

Question 2: "What is the lowest percent correct cut score that would be acceptable, even if no examinee attains that score?" (This value is a participant's estimate of the minimum level of knowledge that should be tolerated, $k_{min.}$)

Question 3: "What is the maximum acceptable failure rate?" (This is the participant's estimate of the highest percentage of failures that could be tolerated, $f_{max.}$)

Question 4: "What is the minimum acceptable failure rate?" (This is the participant's estimate of the lowest percentage of failures that could be tolerated, $f_{min.}$)

Having collected these estimates from each judge the mean value across participants for each response is then calculated. These values are then used in conjunction with actual cumulative outcome data (from the mark distribution) to obtain a cut score.

Figure 2 illustrates a hypothetical application of the Hofstee method in which the mean of participants' judgements about k_{min}, k_{max}, f_{min} and f_{max} were 32.5, 80.0, 12.5 and 90.0, respectively. Ordered pairs of the points (k_{min}, f_{max}) and (k_{max}, f_{min}) are plotted and a straight line drawn between them. Any point on the line defined by these two points is considered to be an acceptable combination of cutoff score and failing rate. The cumulative frequency distribution of the candidates' actual scores on the test (the x-axis being the mark achieved as a percentage of the total raw mark available) is superimposed on the graph. The co-ordinates of the point at

which the two intersect (X in Figure 2) then indicate the compromise percentage correct required (i.e. the cut score for the test) and the corresponding failure rate that would be observed if that cut score were used (here approximately 61% and 42%, respectively).

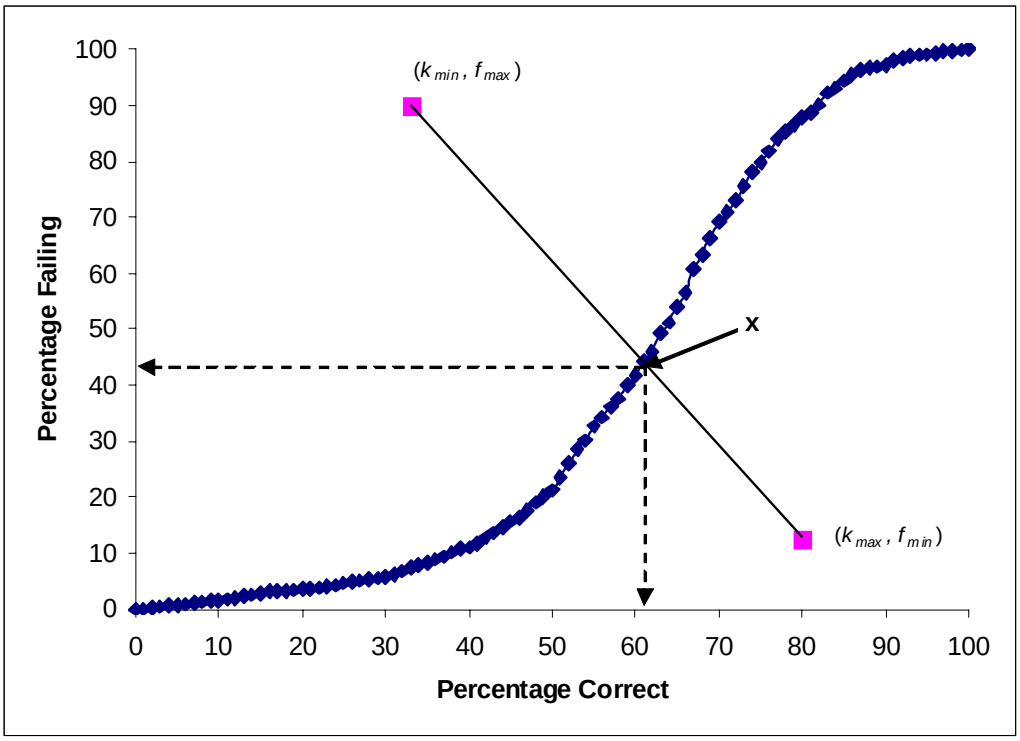

Figure 2: Hofstee's method

This method has the benefit that the data collection is straightforward, it can be applied to any item format and can either be implemented singly or used as supplementary to another method. However, Mills and Melican (1988, page 270) cite their own research, and Cizek (1996, page 27) cites later observations that there is a possibility when using the Hofstee method that the line may not intersect the cumulative frequency distribution, which would result in a failure to identify a cutscore. Another suggested issue is that that, while the expert judgements are on the face of it easy to collect, they lack the detailed approach that is inherent to the absolute standard-setting methods, and can therefore be cursory and lead to estimates that do not fully consider the difficulty of the test.

3.3.2 Beuk's method

Beuk's compromise approach is, in effect, a simplification of Hofstee's procedure, as the judges are asked to respond to (only) two questions:

Question 1: "What should be the minimum level of knowledge to pass this examination?" (This judgement is expressed as the percentage correct for the overall test, *x*).

Question 2: "What passing rate should be expected on this examination?" (This judgement is expressed as a percentage of the examinee population, *y*).

The qualitative answers to these questions are summarised across participants as the mean averages, $\bar{x}$ and $\bar{y}$. Like the Hoftstee approach, these average values are then used in conjunction with actual candidate outcome data to obtain a cut score.

Figure 3 illustrates a hypothetical application of Beuk's method. Assume, for example, that the participants judged that the minimum percentage correct should be 69% and that 89% of candidates should pass, on average. In Figure 3 these points are labelled $\bar{x}$ and $\bar{y}$, respectively. The intersection of points $\bar{x}$ and $\bar{y}$ is then determined and is labelled A in Figure 3. The percentage of candidates that would pass at each of every possible cut score is also superimposed on the graph (based on the mark distribution).

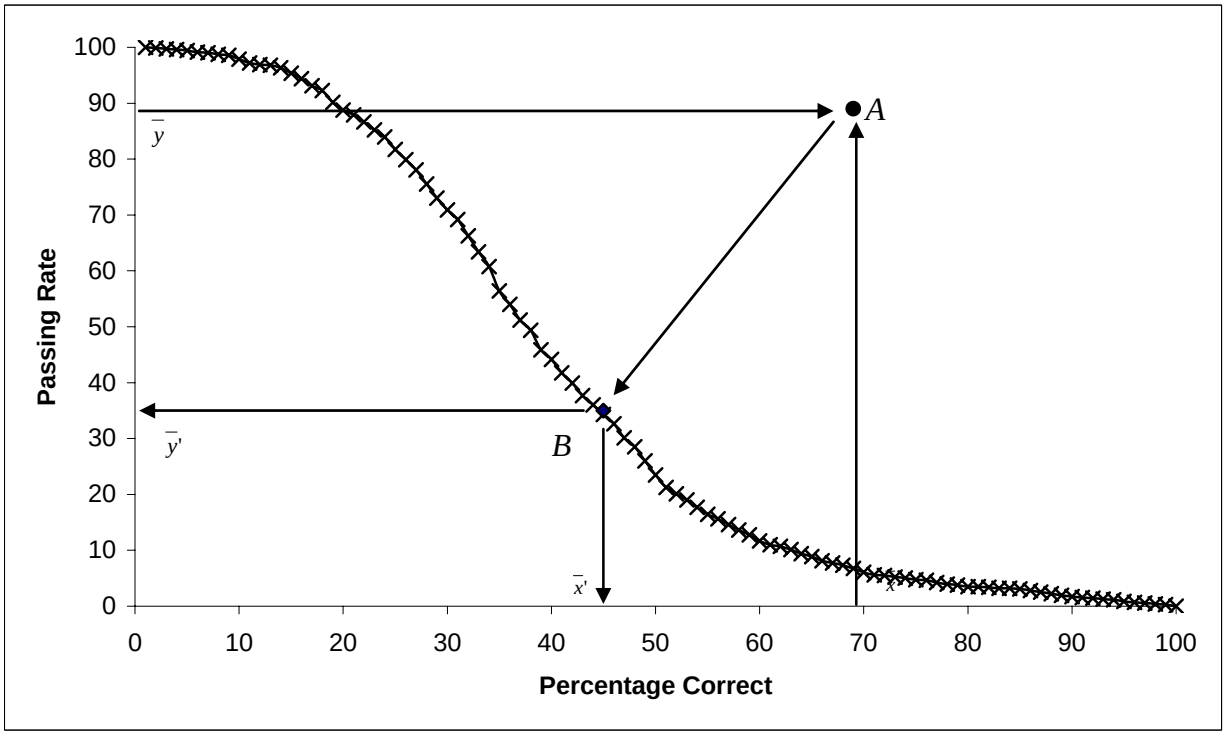

Figure 3: Beuk's method

The standard deviations of participants' answers to the two questions are then calculated (S_x and S_y) and the ratio of the two is used to construct a line with slope equal to this ratio and which, originating at point A, is projected onto the distributional curve. The point at which the line intersects the curve is labelled B in Figure 3. The adjusted values of $\bar{x}$ and $\bar{y}$ are then obtained by projecting point B onto the two axes ($\bar{x}'$ and $\bar{y}'$ in Figure 3). The result determines the 'compromise' (adjusted) percentage correct (i.e. cut score) and the associated passing rate, here 51% and 39%, respectively. (The cut score in raw units is obtained by multiplying the adjusted percentage correct, $\bar{x}'$, by the total number of possible marks available.)

Like Hofstee's method, Beuk's method is easy to apply – both in terms of the judgements which need to be made and the computations involved. A potential drawback suggested by Mills and Melican (1988, page 271) is that judges may have difficulty estimating passing rates if they have not had experience (or are unaware) of the full range of performance of the candidature, and therefore whether it is appropriate to adjust cutoff scores on the basis of variation within the judges is questionable. De Gruijter (1985) attempted to address this issue by incorporating estimates of uncertainty of the individual judges' ratings into the process. However, the method has not been used as frequently as those of Hofstee and Beuk, as it is computationally complex and not as readily understood by judges (Mehrens, 1995). In addition, quantification of the

uncertainty estimates is difficult and appropriate techniques for collecting uncertainty data have not been developed (Mills & Melican, 1988). Further research is therefore needed if De Gruijter's method is to become a viable option.

Interestingly, Hofstee's and Beuk's methods are often suggested as secondary methods to be used alongside other approaches (as opposed to being the prime standard-setting guide for an award). Further, despite their apparent simplicity and time-efficiency, neither method appears to be used as extensively as the other methods covered in this paper. This may be because comparatively little research has been carried out into these methods, in particular very few studies have been published in which results from Beuk's or Hofstee's method are compared with those from other, more traditional, approaches. Also, both were developed in the context of examinations where only a Pass/Fail decision was necessary and there is no record of them being used in a context where more than one cut score is required for a single examination, although it would not appear impossible to extend either method to accommodate this. A practical limitation is that computer programs do not exist to apply the method and analytically derive a cut score, which must therefore be carried out by hand, introducing a potential source of error, but as both methods are fairly simple this would not seem too difficult a problem to overcome.

3.4 The Direct Consensus method

Two disadvantages of the Absolute (item-based) methods is that they are time-consuming, particularly if more than one round of rating is employed and (aside from the arguments about whether or not the judges can actually carry out the task they are asked to perform) the cognitive burden on the judges is high. Another method, the Direct Consensus Method, was therefore developed very recently (in 2004) attempting to address these issues, and is described in detail by Cizek and Bunch (2007).

For the Direct Consensus Method, the original test is reorganised into sections, formed to indicate areas of homogeneous content within the test, and entire sections are reviewed and judged. Participants begin with the first section and are asked to indicate those items they believe the borderline (i.e. just-competent, or minimally-competent) candidate will answer correctly. The participants sum the number of items they have indicated, producing a 'number correct' passing score for that section. This is repeated for each section. The sum of each participant's section scores is then taken as that participant's recommended passing score for the total test. The data are summarised in a format similar to the hypothetical data in Table 4 (adapted from Cizek & Bunch, 2007, page 99), and are provided to the participants for general review.

The hypothetical data in Table 4 indicate the results of the method having been applied to a 46-item test, which has been split into five sections with varying numbers of items per section. Eight participants have indicated the number of items in each section they judge a just-competent candidate should answer correctly. For example, section A comprises nine items and Rater 1 feels that a just-competent candidate should answer seven of these items correctly. The overall mean and standard deviation (s.d.) of all eight participants' recommendations are shown for each section, and for the test as a whole, as the percentage of the number of items in the section that this mean represents. From the Rater sums at the bottom of the table it is clear that Rater 1 would set the lowest recommended cut score (31 items correct out of 46), whereas Raters 3 and 6 would set the highest (35). The final cutscore for the test is the overall mean number of items recommended to be necessarily correct across all the raters, here 33 (although whether the mean should be rounded down, or rounded to the nearest whole mark, where

necessary, is not made clear by Cizek and Bunch). These data would be shown to the participants and a second round of rating could then be implemented, allowing the participants to revise their initial judgements if they wish to do so. The final cut score is then discussed and agreed[17].

Table 4: The Direct Consensus method

	Rater number									
	1	2	3	4	5	6	7	8		
Section (Number of items)	Participants' recommended number of items within a section that the just-competent candidate should answer correctly								Section mean (s.d.)	Percentage of number of items in section
A (9)	7	7	6	7	7	7	7	6	6.75 (0.46)	75.0
B (8)	6	6	7	6	5	7	7	7	6.38 (0.74)	79.7
C (8)	5	6	7	6	6	6	5	6	5.88 (0.64)	73.4
D (10)	7	8	8	7	7	8	7	7	7.38 (0.52)	73.8
E (11)	6	5	7	7	7	7	8	6	6.63 (0.92)	60.2
Rater sums	31	32	35	33	32	35	34	32	33.00 (1.51)	71.7

Cizek and Bunch discuss research indicating the Direct Consensus method to be more time-efficient than the Angoff method and also producing similar cut-scores. However, because the Direct Consensus method has so recently been introduced, very few potential limitations of the method have been explored, neither have possible variations been investigated in any detail. Clearly the method may be more applicable to some tests than others (some tests may comprise items which are less readily divided into sections, for example). Also, as in other standard-setting methods, it would seem sensible for the implications of the potential cut-score to be fed back to the raters during their discussions, so that they are aware of the consequences prior to finalising their recommendations.

4. PROS AND CONS

From AQA's perspective, one of the key features required from any standard-setting method is that it must be readily applicable to situations where more than one grade boundary is to be set on the same test, since it is rare only to have to set one (Pass) boundary on a test paper. Before discussing the pros and cons of the various approaches, it is therefore sensible to eliminate those which fall foul of this criterion. The Nedelsky method is consequently not suitable for our purposes. On balance it is probably sensible to rule out the Compromise methods of Hofstee and Beuk also. At this stage, little comparative work has been carried out to establish the advantages and disadvantages of these relatively new methods. In particular, although it would not seem impossible to adapt these approaches for use in setting multiple grade boundaries on a test, there does not appear currently to be any record of this being attempted. Obviously that does not preclude AQA from attempting to use, say, Beuk's method in awarding a new OT in the absence of other feasible options, but using a well-known and established approach which does not fall foul of the multiple boundary setting criterion is probably more sensible than opting for a new, lesser-known method which has not been tested in that scenario. This leaves the methods of Ebel, Angoff and Jaegar, the Bookmark approach and the Direct Consensus method as potential options for AQA's purposes.

[17] which may, or may not, ultimately be the overall mean suggested by the raters' judgements, as their discussions may lead them to recommend an alternative mark.

4.1 The Absolute methods

Considering the three remaining Absolute methods together, it would be hard to recommend any approach other than one of the Angoff variations. While Ebel's method is adaptable and could be simplified if 'real time' item difficulty data were provided rather than rely on the participants' judgemental categorisations of the difficulty of each item, there remains the potential problem for participants of keeping the two dimensions (difficulty and criticality) distinct and also the logical limitation that in a high stakes examination all the questions should (at least theoretically) be in the upper two categories of importance. Empirically the Ebel method has been shown to have the poorest inter-rater consistency when compared to the Nedelsky and Angoff methods (Colton & Hecht's 1981 paper, cited by Mehrens,1995, page 229; and Cizek, 1996, page 24, amongst others) and ultimately it has not been recommended as a method of choice in comparison to the Angoff-based approaches. Jaegar's method has been introduced comparatively recently and therefore has not received as much scrutiny as the approaches of Ebel, Angoff and Nedelsky. Apart from Berk's (1986) criticism of Jaegar's method not allowing participants to make probability choices other than 0 or 1, research has also suggested it may produce somewhat less reliable standards than the Angoff and Nedelsky approaches (Cross *et al*., 1984). Angoff's method, in comparison, has withstood the test of time and, although not without its critics, remains the most highly regarded, often recommended and widely used of all the standard-setting approaches for multiple-choice tests covered in this paper. Angoff's method and all its variations have become the most thoroughly researched of all the standard setting methods and have been included in many studies with other methods. They are reported as being easy to explain and implement, have been found to produce the most reliable standards compared to other Absolute methods, and thereby offer the best balance between technical adequacy and practicability. As a family of approaches, the Modified Angoff, Extended Angoff and Yes/No method have the powerful advantage that they can be applied to tests comprising mixed formats – multiple choice and constructed response – a flexibility which further sets them above the competitor methods.

4.2 The Direct Consensus method

In contrast to the Angoff approaches, the Direct Consensus method was only conceived in 2004 and is therefore much less well documented and researched. The potential time-efficiency of the method is attractive, as is the lessening of the cognitive burden on the judges - particularly if the early research suggesting the method produces similar cut scores to the Angoff method is reinforced. The time-efficiency of the method may be somewhat reduced if the awarders are informed of the implications of their ratings and allowed to discuss prior to finalisation of the grade boundary, as has been recommended by some reviewers, nevertheless the Direct Consensus method is still likely to take less time overall than, say, any of the Angoff approaches. However, the Direct Consensus method does have the disadvantage that it may be more readily applicable to some tests than others, as in some cases items will group naturally into sections and in others less so. The item grouping is obviously a key part of the method and presumably the allocation of items to sections could affect the final boundary marks awarded, so it raises the question of who should most sensibly carry out the item grouping. (For AQA this would most obviously fall to the Principal Examiner for the test, but this would be an additional responsibility to their existing commitment.) Also, as far as can be ascertained, there is little documented guidance on how the grouping should be carried out (ideal maximum and minimum numbers of items per group, for example). Presumably the advantage over the Angoff Yes/No method, which is, in effect, the same task apart from the lack of item sectioning, is that the awarders are able to consider the relative difficulties of items on a similar topic. Their conclusions as to whether candidates should get each item within the section right or wrong should therefore be more reliable. However, if difficulty of the item is related to its position in

the test the awarders' estimates will be misleading. This point is also relevant to the Bookmark approach and is discussed in more detail in §4.7. Since, because of the need to section items, the approach may be more applicable to some objective test papers than others, and also the method lacks documented testing, it is unlikely to be the approach of choice for AQA, although it does have intuitive appeal.

4.3 The Bookmark method

The Bookmark method, while also relatively new, has become a popular and generally accepted procedure for setting standards on educational tests in the USA, although there is little evidence currently of it being used widely in the UK. The clear advantage of the method is that the judges have to make fewer decisions which are less challenging cognitively since, rather than requiring judges to estimate the probability that borderline candidates will get each item right, the Bookmark method asks judges to establish the most difficult item which a borderline candidate would pass (with a given probability). Also, although the Bookmark method as a standard-setting approach is in its relative infancy, the IRT models and analysis on which it is based are comparatively widely known, which is therefore advantageous to the fidelity of the method and the interpretation of its results. Another positive aspect (generally speaking) is that, while some of the computational aspects of the method are mathematically complex, much of the intensive work is carried out prior to the standard setting itself. Nevertheless, given the basis in IRT, the technicalities of the procedure are more involved and, as already highlighted, there are ongoing questions surrounding important issues inherent to the method which need to be answered to support its wider use.

4.4 Bookmark-based versus Angoff-based approaches

The level of interest in the Bookmark procedure is reflected in the fact that already a few studies have been carried out comparing Bookmark-based approaches with those of Angoff, these methods being arguably the two main rivals. Buckendahl *et al*. (2002) compared the Angoff Yes/No method with a variant of the Bookmark approach, using two rounds of rating to set Pass boundaries for a grade 12 Mathematics assessment comprising entirely dichotomous items. Their Bookmark approach was based on using classical test theory (CTT) item difficulties to order the items rather than the more typical IRT-based mapping procedure, the rationale for using CTT to rank the items (all of which were dichotomous) being that they believed the subject experts, in this case middle school mathematics teachers, would be better able to understand the concepts of classical item statistics as opposed to the ability estimates stemming from IRT-based methodology. (The correlation between the rank ordering of items according to the IRT and CTT mapping strategies was very high, -0.98, thus the ordering of items was not greatly affected by the use of the simpler CTT approach.) Another departure from the standard Bookmark procedure involved the cognitive task required of participants when placing their bookmarks. Participants were asked to conceptualise a barely proficient student (BPS) they had taught in the past. Keeping that student in mind, they were then told to, "start with the least difficult item and move through the booklet until they found the place where the BPS would probably get all the items to that point correct and probably get all the items beyond that point incorrect" (Buckendahl *et al*., page 256). At that point in the booklet the participants placed their bookmark and thus a probability of success (e.g. 0.67) was not defined. Having had the opportunity to revise their initial bookmarks in a second round of rating the recommended cut score was calculated by summing, for each participant, the number of items up to the bookmark and then averaging (mean or median) those values across the participants. The differences in the Bookmarking approach obviously cast some doubt on the generalizability of the results from this study. Nevertheless, the findings show some evidence of consistency of final results between the methods, the final cut scores being only two marks different, in this

case the Angoff approach ultimately yielding a lower boundary than that from the Bookmark method. Between rounds, the cut score from the Angoff method dropped by 1.5 marks, although the standard deviation of the judges' estimates increased (which is somewhat surprising) by just over three marks. In contrast, the cut score from the Bookmark method increased by two marks (and the standard deviation decreased), which is not too dissimilar to Yin and Sconing's findings in 2007 (see later text). Overall the authors suggested that using CTT as the basis of the Bookmarking procedure rather than IRT methods may be more practical, particularly when sample sizes are too small to calibrate the IRT-based parameters or when IRT methods are not feasible because of, for example, a lack of software availability.

Wang (2003) compared the results of using a Rasch IRT model in his item-mapping (Bookmark variant) approach with those obtained using the traditional Angoff method[18], based on two rounds of rating in each case. Unlike Buckendahl's findings, the final cut scores from the item-mapping and Angoff approaches differed by between eight and ten marks, the item-mapping approach producing lower cut scores, and consequently higher potential pass rates, than those from the Angoff procedure. Nevertheless, in Wang's study, the judges were apparently satisfied with their predicted pass rates (from the item-mapping), believing their cut scores met their conceptual understanding of the passing standard. His results indicated that judges provide more consistent ratings when using the item-mapping method compared to the Angoff method and also, although rater agreement was very high in both methods, it was consistently higher in the item-mapping method. However, there are flaws in the study design which cast some doubt on the validity of the outcomes: three different groups of judges were used to set cut scores on four licensing examinations (not related to education), each set of judges using both methods, i.e. Angoff followed by item-mapping, or vice versa. It is clear from the results that the order of implementation affected the outcomes, the rater agreement between the two methods being higher for the exams where the Angoff method was used first to obtain a cut score. The presentation order appears therefore to have contributed to the discrepancies in the estimates of the average consensus amongst the judges.

Karantonis and Sireci (2006) cite a paper by Yin and Schultz (2005), presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada, which compared cut scores and cut score variability from Angoff-based and Bookmark-based procedures in standard setting. The Bookmark-based procedure used was a recent development of the method, called Mapmark (Schultz & Mitzel, 2005; Schultz, Lee & Mullen, 2005). Essentially the first round of the Mapmark approach is the same as the Bookmark method, but in round two the use of domains and domain-score feedback is introduced, in tabular and graphical form, domains being areas of knowledge, skills and abilities (KSAs) into which items are pre-clustered[19]. The Mapmark procedure therefore facilitates discussion about what KSAs are being measured by the items, as well as the KSAs students who 'master' these items possess. Notwithstanding the variations between the traditional Bookmark and Mapmark approaches, Yin and Schultz's results indicated a tendency for the Bookmark-based cut scores to be lower than those from the Angoff-based approach.

More recently, a study by Yin and Sconing (2007) also compared Angoff-based and Bookmark-based (again Mapmark) methods, examining the variability of cut scores resulting from both

[18] i.e. requiring judges to estimate, for each item, the probability that a minimally competent candidate will answer the item correctly.

[19] Very loosely, the basic concept is therefore akin to that behind the Direct Consensus method, in that in the Mapmark approach the panellists are provided with information about the relative difficulty of items within item subgroups (domains), in addition to the overall item difficulties used to structure the ordered item booklet. For example, in the mathematics test on which the study by Schultz, Lee and Mullen is based, the items could be grouped into five content domains: Number Sense, Measurement, Geometry, Data Analysis and Algebra.

approaches. Panellists were divided randomly into two groups, that either used the Angoff-based or Bookmark-based procedure to set standards for three grades (Basic, Proficient and Advanced) on a National Assessment in Educational Progress grade 12 Mathematics examination using four rounds of rating in each case. Over rounds, the cut scores for the two methods tended to converge and the standard deviation of the judges' estimates became smaller, although the cut score estimates from the Angoff-method reduced across the rounds (by between seven and eleven marks[20]), whereas those from the Bookmark-based approach increased (by between one and five marks), i.e. the panellists made less adjustment to their initial ratings when using the Bookmark-based method than when using the Angoff method. Ultimately the cut scores resulting from the Bookmark-based method were lower than that of the Angoff method for the Basic and Proficient grades (by four marks and six marks, respectively) and were eight marks higher for Advanced grade. Considering the standard error (SEM) of each cut score, the SEMs were relatively small for both methods but, whereas the SEMs for the Bookmark-based cut scores were much larger at round 1 than round 3, those for the Angoff method showed less of a difference. Overall Yin and Sconing's results suggested that although the Bookmark-based cut scores were more variable at the first round they became similarly stable as Angoff-based cut scores over rounds.

4.5 Using IRT data to assist judges in making Angoff item judgements

4.5.1 Presenting item performance information for every total mark on the test

The problem with feeding back facility information to awarders using the Angoff or Bookmark approach in meetings is that such data relate to the average performance of candidates rather than the performance of only those candidates near the grade boundaries. Both Fearnley (2003) and MacCann and Stanley (2006) have suggested, in the context of the Angoff method, that IRT data could give awarders information about performances of candidates on all items for each total score on the test paper, i.e. on all possible grade boundary marks. A table (such as Table 5 – adapted from MacCann & Stanley) could be prepared showing how candidates on each total score would be expected to perform on each item. (For the purposes of brevity, only a selection of total scores is shown in the table and also only the first four items on the test.). For example, according to the (in this case) Rasch IRT model, of the students achieving 45(/50) on the test overall, approximately 95% would be expected to get item 1 correct but only 72% would be expected to be correct on item 2. In addition to providing information of candidates' expected performance on an item, this also provides a way of comparing the difficulty of the items at different ability (total mark) levels: item 2 is generally a more difficult item than item 1, for example. Comparing item 1 with item 3 indicates little difference in the probabilities of a correct response on a total mark of 47 (97% versus 94%) but the difference increases as the total mark (ability level) drops.

MacCann and Stanley suggest presenting the data, such as those in Table 5, to the judges *before* they carry out their Angoff estimates. They suggest the judges should, independently, scan the item probabilities for item 1 and encircle the probability that best matches their own Angoff rating of that item. They would then repeat the process for item 2, and following items. If the encircled probabilities tended to fall on the same row, they could stop when they were satisfied that the row reflected the standard they are thinking of. Some judges may need to work through only a small number of items before settling on a row, whereas others may need to cover most (or all) of the items. Once a row is tentatively chosen, the remaining items could

[20] out of approximately 108, the 180 (one-mark) items in the test were divided into two sub-pools, A and B, each comprising approximately 60% of the items in the assessment.

be dealt with more quickly, as a confirmatory check (probably by the group of judges as a whole), rather than performing Angoff ratings in isolation.

For example, the judges' (average) Angoff ratings for the first four items are emboldened in Table 5. The preferred cut point appears to be a score between 41 and 43 (and ultimately these three marks could therefore be discussed further, alongside the cumulative percentage outcome on the marks and other statistical information to determine the final boundary to be recommended). Also it is clear which Angoff ratings are aberrant, as for item 2 – this is a relatively difficult item, for which the judges have underestimated the difficulty experienced by the candidates. Using this approach it is easy for the judges to see these discrepancies and to amend them. Since this procedure presents actual (modelled) item data upfront to the judges they are able to use these to inform their Angoff judgements, scanning across a range of items to see the probabilities the (Rasch) IRT model has assigned and comparing these with the estimates they would have provided. They are able, potentially, to 'home in' on a particular mark (i.e. a particular row in the table) early on, rather than following the standard Angoff process of making sequential item judgements necessarily across the whole paper. Thus the approach benefits from the ease with which aberrant judgements can be detected and also the likely time saving on the standard Angoff process.

Table 5: Probabilities of success, as estimated from a Rasch model, for students at selected total marks for the first four items on an OT, along with the cumulative probability of candidates on each mark

		Probability of success (Rasch model estimates)			
Total mark (/50)	Cumulative %	Item 1	Item 2	Item 3	Item 4
:					
48	99.5	.98	.88	.96	.98
47	99.1	.97	**.83**	.94	.97
46	98.3	.97	.77	.92	.96
45	97.1	.95	.72	.90	.95
44	95.7	.94	.67	.87	.93
43	94.1	.93	.62	.85	**.92**
42	92.3	.92	.57	**.82**	.90
41	90.2	**.90**	.53	.79	.89
40	88.1	.89	.49	.77	.87
39	85.9	.87	.45	.74	.85
38	83.5	.85	.41	.71	.83
:					

Fearnley (2003) also suggested presenting data such as those in Table 5 to the awarders, but he proposed these data be shown at the second stage of the Angoff process. Fearnley proposed redefining stage two to require the judges to select the most appropriate set of item success rates that represent the grade boundary. If, for example, the awarders' Angoff estimates after stage one suggested a boundary mark of 11, they could review the sets of item success rates on marks on and around 11 to establish their final boundary. Thus unlike the standard multi-stage Angoff approach, after stage one, no further estimation of individual item probabilities would be required from the judges. Fearnley suggested that the awarders could benefit in the long term from the feedback of the precise values they had been trying to estimate and that their skill in estimating item difficulty could therefore be improved for future examination series.

It is worth noting that although both Fearnley's and MacCann and Stanley's suggestions are based on presenting tabled IRT estimates, CTT based facilities (i.e. the actual proportion correct for each item on each total mark) could alternatively be provided (see later discussion in §5).

4.5.2 Feeding back to the judges on the accuracy of their probability ratings

MacCann and Stanley put forward a further suggestion as to how (Rasch) IRT may be used to inform judges' Angoff probability estimates. Having made their standard Angoff judgements in stage one of the process, the judges could be presented with information on their own (average) estimates on each item, along with the Rasch item difficulty and the probabilities of success on the items as estimated by the Rasch model (see Table 6 - again for brevity only ten items are included, but the table would be prepared and presented for all items in the test). The tabled data could be presented in item order (as in Table 6) or according to the differences between the judges' and Rasch estimated probabilities of success; the latter would clearly indicate the items where the judges' Angoff estimates were most discrepant from those of the Rasch model. In Table 6, item 16 is highlighted as a particularly difficult item, for which the Rasch estimated probability of success was only 0.55. However, the judges expected the candidates to perform well on the item, with an average probability of success rating of 0.90. Being pointed to this discrepancy would enable the judges to re-examine this item, attempt to determine why it proved so difficult and, if they wish, to readjust their ratings.

Table 6: Comparison of judges' Angoff estimates and Rasch estimates of probability of success

		Probability of correct response		
Item	Rasch item difficulty	Judges' Angoff estimate	Rasch estimate	Difference (Judges – Rasch)
10	-0.327	0.99	0.95	0.04
11	2.231	0.73	0.60	0.13
12	1.702	0.85	0.72	0.13
13	1.007	0.85	0.84	0.10
14	-0.687	0.91	0.96	-0.05
15	0.772	0.87	0.86	0.01
16	**2.442**	**0.90**	**0.55**	**0.35**
17	1.561	0.73	0.74	-0.10
18	2.066	0.78	0.64	0.14
19	-1.105	0.99	0.98	0.01

While not reducing the time expended by the Angoff method, the direct comparison of judges' estimates and Rasch estimates would focus the judges' discussions on the discrepant items and thus could clearly be a helpful tool in establishing the final grade boundary to recommend.

4.6 AQA's experience of the Bookmarking method

While the Bookmark method is attractive in that it would pose a simpler cognitive task to the awarders than the traditional Angoff approach and could ultimately be less time-consuming once the use of the method was established, trials of the Bookmarking method carried out a few years ago in AQA, following an earlier review of the judgement-based approaches to setting boundary marks, were not promising (Fowles, 2003; 2004 & 2005). The first trial was based on candidates' responses to the objective test items in GCSE General Studies Paper 1 Section A

from June 2003 and applied IRT one- and two- parameter models to the test item data for the total candidature (1300 candidates across Foundation and Higher tiers)[21]. The second trial was based on candidates' responses to one of the GCSE Science Double Award Modular module tests in November 2004 and applied a Rasch one-parameter model to test item data for a random sample of 5000 Higher and 5000 Foundation tier candidates to produce a single item booklet (the Rasch one-parameter model being used in the second trial in response to the recommendation from the November 2003 Research Committee meeting, mentioned earlier (§3.2.1)). In each trial, when setting their bookmarks for each judgemental grade (C, A and F, in that order), the item writers and examiners were asked to follow the traditional Bookmark approach, i.e. to think of a borderline candidate and identify the last item they would expect that candidate to answer correctly, with correctness defined as having a 2 in 3 (67%) chance of success[22]. The Bookmark positions provided by each judge in each item booklet were then translated into boundary marks to compare with those established in the awards. In the first trial, as a follow-up exercise the judges were also asked to provide traditional Angoff estimates for grade C in both tiers, to allow some comparison of the boundary mark suggested by the two methods.

In both trials there was wide variation in the locations of the Bookmark placements between the judges. Although these differences generally reduced when translated into boundary marks, it suggested that Bookmark method judgements could not be taken for use in boundary setting without substantial amounts of training and practice. In contrast, the first trial judges' estimates from the Angoff procedure showed reasonable consensus[23] and, while the potential boundary marks from the Angoff and two-parameter model Bookmark approach were not too dissimilar, there was a much greater differential between the Angoff and one-parameter model results, the one-parameter model estimates being likely to converge to a lower boundary than those suggested from the Angoff process. The judges' comments after the first trial revealed insecurity and a lack of confidence in carrying out the Bookmark task, which could perhaps be overcome with experience. However, in both trials the participants reported problems with some items on accepting that they were indeed more difficult for candidates than the preceding items. Thus the potential advantage of the method, that judges do not have to consider items' relative difficulties, was not fully realised and in some cases became an issue in carrying out the task. There was also evidence that judges had found it difficult to apply the concept of a 67% probability of success, highlighting the fact that although the Bookmark method removes the necessity for judges to estimate the specific probability of success on every item, it still requires them to be able to estimate a borderline candidate's probability of success on any given item. While this might be expected to improve as a result of experience, training and feedback from item analyses, the same is also true for the Angoff method, spawning the question of whether much ultimately is gained by using the Bookmark approach as it stands, particularly if questions are raised regarding the ordering of the items in the item booklet. From the functional perspective, a major issue encountered during the AQA Bookmarking trials was the time consuming preparation and calculations necessitated by the method. For the one-off exercise, the item booklet was constructed by cutting and pasting from the original examination paper, which was an extremely lengthy, fiddly process, and obviously if the method were to be actively used a more polished, less burdensome approach to developing the booklet would have to be found. Even so, the detailed computational preparation prior to the award (and, potentially, during the award, when calculating the final grade boundary) was also lengthy. Overall

[21] Two similar, but not identical, item booklets were therefore produced and received Bookmark judgements in the first trial.
[22] In practice, to help the awarders' interpretation it was suggested they think of three borderline candidates and consider whether they would expect two candidates to get the item right and one to get it wrong.
[23] Bear in mind that the judges had no known previous experience of either method.

therefore the experience of the process was that it was operationally long-winded and awkward, as well as not being a great success judgementally.

As a consequence of the Bookmark trial experience, Fowles (2005) suggested an alternative approach, drawing on Fearnley's (2003) proposal of using IRT estimates to inform the Angoff process (discussed in §4.5.1). Fowles' alternative Bookmarking approach would mean redefining the task. Instead of working through the item booklet and selecting a single Bookmark item, the judges would be presented with a table of data (such as Table 7) showing how candidates on each total score would be expected to perform on each item according to the IRT (or CTT) analysis. (The data would be prepared covering all total marks in the test and all items, but for the sake of brevity only a subset of items and total marks are shown here.) For each total mark, the items would be listed in order of difficulty and the item most closely associated with the 67% success criterion identified (for example, these items are emboldened in Table 7). The '67% items' would therefore be highlighted upfront to the judges. The judges' task would be to identify a likely boundary mark[24] and decide if the item identified with a probability of success closest to 67% for that mark is the most appropriate item to represent performance at that grade boundary. If not, the judges would review the alternative 67% items on the adjacent marks. For example, if item 11 in Table 7 was considered to represent appropriate performance at the particular grade (grade A, say) the recommended boundary mark would be 30. Although the data would sensibly be prepared to cover the full mark range, cut down versions of the full table, showing only specific ranges of total marks need be presented to the awarders to focus their attention on the marks on and (say, in a five mark range) around the likely grade boundaries. In this alternative method, there would be no need to produce an ordered item booklet - data of the kind presented in Table 7 would be the stimulus material, along with the question papers and the standard item facilities for the test overall - and the awarders' judgements on the 67% items would be the guide to the final boundary mark. Clearly training would be required to help awarders interpret data such as those in Table 7 and time would have to be allowed in the award for them to discuss and reconcile their views. It should also be noted that the judges would still need to grasp the meaning of '67% probability of success' to carry out this task, just as when identifying the Bookmark item in the ordered item booklet. Although considered an interesting possibility at the June 2005 Research Committee meeting, to date this approach has not been explored further.

Table 7: Example probabilities of success for candidates on a selection of marks on an objective test with a maximum mark of 36 (and comprising 36 items, not all shown in the table)

	Total mark on the objective test								
Item	32	31	30	27	26	25	21	20	19
5	0.97	0.96	0.91	0.86	0.85	0.84	0.73	0.69	**0.65**
8	0.95	0.92	0.86	0.84	0.83	0.77	0.71	**0.67**	0.64
3	0.93	0.91	0.84	0.80	0.78	0.76	**0.66**	0.62	0.60
7	0.89	0.86	0.81	0.79	0.75	0.72	0.65	0.61	0.59
2	0.86	0.84	0.75	0.76	0.73	0.69	0.62	0.58	0.56
1	0.83	0.81	0.73	0.73	0.72	**0.67**	0.60	0.57	0.53
4	0.79	0.75	0.72	0.72	**0.67**	0.62	0.57	0.55	0.50
6	0.76	0.73	0.69	**0.68**	0.65	0.61	0.49	0.43	0.41
11	0.73	0.72	**0.66**	0.65	0.62	0.58	0.46	0.40	0.37
9	0.72	0.70	0.62	0.62	0.59	0.56	0.44	0.39	0.36

[24] this could be the SRB, for example, which typically might be the boundary mark suggested by the unit-level prediction.

4.7 Is preserving the item sequence an issue?

A further question which could affect the decision on which method to use to award an objective test is whether it is important to preserve the sequence of items as originally presented on the paper. (This discussion therefore relates particularly to the Direct Consensus method and the Bookmark method.) Whether candidate performance is affected by the order of the items on multiple-choice tests has been the subject of much research, amidst speculation that the proportion of examinees who correctly answer a test item may be influenced by the difficulty of the preceding item or, more specifically, that the increase in anxiety level brought about by exposure to a hard question interferes with examinee's ability to deal with the next item. Clearly this is a difficult question to answer conclusively as it would be very difficult to be able to generalise results to all students and all testing situations. If a sequencing effect did exist it would probably exist to varying degrees depending on the distribution of item difficulties, the perceived importance of the test, etc. Although there are exceptions (Balch, 1989; Vos & Kuiper, 2003; for example) most studies fail to find a significant effect of the sequencing of items on overall candidate performance in the test (Huck & Bowers, 1972; Laffitte, Jr., 1984; Newman, Kundert, Lane & Bull, 1988; Perlini, Lind & Zumbo, 1998; and Neely, Springston & McCann, 1994). Generally therefore we should probably not be concerned about whether the item's position has affected the students' perceived difficulty of that item. Consequently, is it acceptable to alter the item ordering in order to obtain awarders' estimates of the item difficulty for the purposes of grade boundary setting? On the basis that the item's position is unlikely to have affected the difficulty of that item for the candidate, if an awarder considers the item's difficulty in a different sequence to the item's original position on the paper it should not matter. However, if there is any concern that position may influence the item difficulty then using an awarding method which alters the item order, i.e. the context of the question, for the candidate and the awarders would be questionable in terms of its validity. The awarders would be considering the items in a different sequence to the way they appeared to the candidates taking the test, thus their estimation of the difficulty of each item may be different to its difficulty in the context of the paper.

5. DISCUSSION

The two approaches which appear to be the most sensible options available to AQA for the awarding of its objective test papers seem to be the Angoff and Bookmark methods. While the Direct Consensus method and Beuk's method are also potentially attractive options they are likely to be less applicable to AQA's needs in general than Bookmarking and Angoff approaches, the Direct Consensus method requiring items to group into logical sections and Beuk's method not having been applied to setting multiple boundaries on a test.

For AQA, the Angoff approach has the drawback suffered by all the Absolute methods - predicting item performance is not a task that AQA's awarders are usually asked to perform. Awarding committees generally evaluate individuals' performance on overall papers, albeit on occasion with specific reference to individual questions which have been flagged by the Principal Examiner to be particularly predictive of (say) grade A achievement. Specific predictions of item-level (i.e. question-level) performance are not part of the normal standard-setting process. The traditional Angoff method has had mixed success in the awarding of AQA's GCE objective test components, although this is partly due to the structure of the units in which it is employed and also the focus of the awarding process overall. The aim in awarding is to carry forward unit standards as far as possible while primarily maintaining the standard of the overall subject, therefore predicted outcomes (adjusted for candidates' prior GCSE achievement) are provided at unit- and subject-level for all AQA's GCE awards. As explained in

§2.2, in relation to GCE Chemistry, since the OT combines with a coursework or a practical component (for which the boundaries would be expected to carry forward year on year) to form optional alternatives for the unit, there is limited scope for where the grade boundaries can be set for the OT if the unit-level standards and comparative outcomes between the two optional units, are to be maintained. Thus the unit-level predictions will guide the setting of the OT boundaries more strongly than the awarders' Angoff judgements and there is therefore little point in carrying out more than one round of rating[25]. In GCE Economics and General Studies A the situation is similar in that the awarders choose to set standards on the written component of the unit first, following the standard scrutiny procedures. Having done this the scope for the OT boundaries is, in essence, guided by the unit-level prediction. GCE Physics A is the exception in that the awarders prefer to carry out two rounds of Angoff judgements to establish the OT boundaries and then allow the written paper boundaries to be guided by the unit-level prediction. Clearly therefore the Angoff approach has worked suitably well in this subject for the awarders to use it to direct the setting of the OT boundaries so firmly (as opposed to focussing primarily on the written paper). As with any standard-setting method, the Angoff approach benefits from practice and experience, and it is reassuring that, even in GCE Economics where the weight of the awarders' discussions is focussed on their written paper judgements rather than those from the OT, the correlations between the awarders' individual estimates of the percentage correct responses for each item and the actual facility indices have been seen to improve (Meyer, 2003a; 2003b). This is important in the context of the possible use of the Angoff method in awarding the new Diploma OT examinations. In the Diploma, the objective tests are (as far as is known) all single component units, thus the complications encountered in the GCE scenario with two component units will not be relevant and the focus will be entirely on the OT. If an Angoff method were used for awarding series on series, improving the accuracy of the awarders' estimates would be a prime concern, which could potentially be approached by using Fearnley's or MacCann and Stanley's approach (discussed in §4.5.1 and §4.5.2)[26]. The recent concept put forward by Béguin *et al*. of adjusting the awarders' current estimates according to previous leniency or severity is an interesting idea, but was found by the authors not to be effective (see §3.1.3(d)).

Although the criticisms of the Angoff-method have been strongly refuted, the research indicating many judges are unable to accurately predict item-level performance is hard to ignore. This does not come as a surprise given AQA's own recent research showing that, while awarders are able to make broad judgements in identifying standards, they are unable to make fine distinctions between candidates' performances in terms of grade worthiness (Baird & Dhillon, 2005). Following this research, and also given ever-increasing time-pressures on awarding in general, attempts are being made to streamline the awarding process by establishing a more appropriate balance between the validity of the qualitative judgements and the fine-grained detail of statistical data. Recent trials of alternative approaches to awarding written papers, termed the Confirmation and Zone of Uncertainty Methods (Stringer, 2007) focussed the awarders' judgements respectively on establishing whether scripts exactly on the statistically recommended boundary (SRB) appear to be of an appropriate standard (i.e. requiring the awarding committee to scrutinise only scripts on the SRB), and alternatively identifying an area of the mark scale in which they believed the boundary to lie (thus establishing the range, but not then being expected to establish a preferred mark within that range). Most recently, in the June 2008 awarding series a 'three-mark range approach', which focuses the awarders on the SRB and the two marks immediately either side, was trialled in the awards of written components in

[25] which at least incorporates an element of professional judgement in the standard setting process for the OT, as preferred by Ofqual.
[26] Using IRT to link standards across series could be an alternative, which would obviate the need for using an Angoff-type approach *ad infinitum*.

stable GCE specifications by most Awarding Bodies. The evaluation of this method is currently being prepared by STAG for the Office of the Qualifications and Examinations Regulator (Ofqual), and it is intended to use the same approach to award the legacy GCE A2 units in January and June 2009. Thus, in the current context of research and development in the awarding of written papers, and given the known limitations of awarders' judgements when based on the total marks achieved on an overall paper, introducing more widely an awarding procedure for objective tests which requires individuals to make fine-grained *item-level* judgements seems something of a backward step, even if that method is well-established and generally highly regarded. With that in mind, the Yes/No variation of Angoff method may be preferential to the other Angoff approaches, as awarders would then simply have to indicate whether a borderline candidate would answer each item correctly, which is a far broader task than required by the traditional Angoff approach currently employed in the GCE scenario. The judgemental thinking required in the Yes/No method is closer to the Bookmark approach when implemented with a 0.50 probability of success, but to date a comparative study of these two variations of the methods does not appear to have been carried out.

While the Angoff approach is not without its problems in terms of the cognitive pressure it places on awarders, it is inherently simple to apply. In contrast, although the Bookmark procedure attempts to remove much of the pressure on the awarders in terms of decision-making, even if Fowles' (2005) alternative to the Bookmark procedure, discussed in §4.6, was implemented (unless CTT-based), the IRT modelling which supports the Bookmark method remains mathematically complex and the preparation lengthy. While much of the analysis necessary would be carried out prior to the awarding meeting, the data would still need to be interpreted for the awarders in the meeting, which has implications for the selection of Support Officers who could reasonably be expected to service the award. Further, within the meeting, ultimately a grade boundary would have to be calculated from the awarders' individual judgements, thus some form of awarding software incorporating IRT Bookmarking methods would need to be developed to be able to feed back the implications of the awarders' recommendations in 'real time'. Very few Senior Research Officers (SROs) at present have detailed knowledge of IRT modelling and analysis, thus to maximise the potential network of Officers within the Research and Policy Analysis Department (RPA) who could be asked either to run analyses prior to an award, and/or offer backup advice and guidance to an award even if not attending in person, some form of training on IRT for Research Officers would sensibly be required. Nevertheless, the demands for Support Officers during the awarding series are such that there would be no guarantee of a SRO being available to run IRT analyses for an additional award they are not supporting personally, let alone act as a dedicated Support Officer to an award using Bookmarking to set OT grade boundaries. The Support Officer is instead very likely to be from a subject department, or a senior member of staff. These latter individuals would also need to be able to grasp the basics of the approach and be able to interpret the results of the awarders' judgements, even if not being aware of the details of the model, to be able confidently to service the award. A non-RPA Officer could not be expected to run detailed IRT analyses when preparing for an award, nor indeed for the standard Bookmark approach to prepare an ordered item booklet (OIB). Thus if a Bookmarking approach was to be used in earnest for OT awarding, bespoke software would have to be created[27] to establish the item facilities, create the OIB and establish the final grade boundary from the awarders' recommendations, thereby alleviating much of the pressure in the pre-award preparation.

[27] Possibly adapted from the existing Computer Assisted Design, Analysis and Testing System (CADATS), which has been developed internally by Qingping He (SRO) for creating items for an item bank, test construction and analysis.

In the current climate of ever-increasing numbers of awarding meetings, any prospective limitation on the individuals able to support awards using a Bookmark approach in the boundary setting process is a particularly significant consideration against the use of the method and there seems little point in using an approach that is more complex than it needs to be. Following the suggestion by Buckendahl *et al*. (2002) of using CTT, rather than IRT, to order the items may therefore deserve serious consideration. As long as there is no need to make use of the benefits of IRT (in terms of linking common items across tests, for example), and also, in the case of Fowles' (2005) variant to the Bookmarking approach, as long as the observed data showed no irregularities between different total score groups (which would make the awarders' task unworkable with CTT data) using CTT rather than IRT (Rasch) may be preferable as standard facilities are more generally understood by subject staff and senior examiners. Although using CTT is not without its problems (discussed in §3.2.1), the issue of candidates not reaching items need not be worrisome if steps are taken to ensure that suitable time is allowed for the test (which is the intention in the development of AQA OTs). Further, as long as no multi-mark items are included in the test (which is the case in most OTs) the ordering of CTT facilities should be the same as the item difficulties obtained from a Rasch model. Some software development would still be necessary but operationally a CTT, rather than IRT, Bookmarking approach may be more practical in the long-term.

Another factor of particular importance from AQA's perspective is the number of awarders required to establish secure results from any individual awarding method. AQA's awarding committees are required to comprise at least four members, and usually the committee numbers more than this, often involving up to eight members of the senior examining team, and sometimes more. Despite the abundance of references investigating, for example, Angoff-methods in general, the question of how many subject matter experts should be involved in the grade boundary decisions when using the Angoff method to ensure adequately dependable ratings is scarcely documented. Hurtz and Hertz's research in 1999 led them to conclude that generally between ten to fifteen raters is an optimal target when implementing the Angoff method, but that a smaller panel can also sometimes be sufficient – a very small panel can produce stable cut scores. Earlier work (which they cite on page 886) carried out by Livingston and Zieky in 1982, suggested that as few as five raters can be used, although more if possible. Thus, from AQA's perspective it appears our awarding committees usually comprise sufficient numbers to produce reasonably reliable estimates when using the Angoff approach at least. Further, Yin and Sconing's recent study (2007) of the Angoff and Bookmark-based methods only involved between five and six raters per group and, as far as can be seen, this did not appear to be detrimental to the results from either method.

While all the methods discussed in this paper are, depending on one's point of view, equally valid in terms of *setting* standards on an OT, fundamental to the awarding process is the *maintenance* of standards. As discussed earlier, over the past decade it has become customary within AQA to approach the maintenance of standards via the use of predicted outcomes[28] for each subject (and unit). There is no reason why the outcome for an OT unit cannot be predicted in the same way as for a written unit or a portfolio unit and indeed, maintaining standards on the current GCSE Science OT modules is primarily achieved via the use of this approach. In addition, IRT is a particularly useful tool in GCSE Science for ensuring that grade C standards are maintained across the tiers of each OT module. (Thus, for the reasons discussed earlier, it may be preferable to focus the use of IRT on standards linkage - across tiers or, potentially, across series, for example - rather than widening its application via, say, a Bookmarking approach.) However, Ofqual is not entirely content with the GCSE Science

[28] whether based on AQA-only data, or inter-board data, and adjusting for prior candidate achievement

awarding procedures because of the lack of professional judgement involved[29]. Therefore for future OTs, an approach which maintains standards via the use of, say, predictions (and IRT), but also incorporates some form of expert judgement in the process is preferable, at least from the point of view of the Regulators. Nevertheless, the problem remains that expert judgements on OT items are likely to be just as, if not more, unreliable than those on written papers. If we were confident that awarders could competently carry out the task being asked of them, for example estimating the proportion of candidates who will get an item right, we could be relatively secure in their resultant boundary and that standards were being maintained. If, say, the items on the test were harder one year compared to the previous year, awarders would adjust their estimates of the proportions of candidates achieving a correct answer accordingly and the overall boundary would therefore reflect the changed difficulty of the paper – however, as we know, it is very unlikely that the degree of adjustment by the awarders will be correct (if it happens at all!). Therefore we are likely to be more secure in *maintaining* standards on an OT by using a statistical approach, such as predicted outcomes, to focus the awarders on an initial mark, the SRB, but then incorporating in the award discussion of, say, item data on the SRB and the immediately adjacent marks either side to involve the element of examiner judgement (akin to the approach for written papers). Why ask the awarding committee to put effort into making estimates at item-level each series when their doing so will not necessarily maintain the standard they set the previous year? Indeed, since accurate data on what they are being asked to estimate can be provided at (or even prior to) the award, why should they be requested to carry out what then becomes a futile and somewhat redundant exercise? Consequently it would seem very sensible to incorporate the suggestions made by Fearnley (2003), McCann and Stanley (2006) and Fowles (2005) of presenting item facilities on each total mark (or restricted ranges of marks) to the awarders – at least to facilitate their discussions, even if not replacing a 'full blown' Angoff or Bookmark approach completely at all series.

Despite this, in an initial series, the amount of professional judgement involved in setting OT boundaries will necessarily be greater to ensure that standards are set appropriately, particularly as any predicted outcome for a unit (whether that be a written unit, portfolio or OT) in the first series of awarding will inevitably be highly tenuous[30]. Thus this still leaves the outstanding question of, on what approach should the awarders' independent judgements for objective tests be based? While the approach would ideally be applied generally in OT units across more than one qualification type, it helps to focus on the imminent needs of the Diploma OT awards. The awarders for the new Diploma units are likely to be new to the task. A judgemental process that is relatively simple, and easy to explain and implement is therefore very desirable. On the other hand, a technically detailed approach which removes much of the cognitive challenge to the awarders would also be attractive. Although there are relatively few OT units in the Diploma, the written and internally assessed units are plentiful, thus speed of the method is also important. In summary, AQA ideally requires an awarding procedure for objective tests which (list not exhaustive):

i) incorporates some form of awarder judgement;
ii) makes acceptable cognitive demands of the awarders;
iii) can readily be used to set more than one grade boundary on a test;
iv) is relatively easy to explain, understand and implement;
v) is time efficient.

As previously noted, it is likely that there will be an SRB (of varying security, depending on the series) provided to the award to compare alongside any mark suggested by the awarders'

[29] Even though it could be argued that the key professional input takes place instead at the paper-setting stage.
[30] although no more so than the awarders' judgements, necessarily!

judgemental estimations. Also there could sensibly be variations in the approach used for the initial standard setting compared to that employed in future series to maintain standards. Bearing both of these points and the preceding discussion in mind, the Angoff and Bookmarking methods each offer options which satisfy most of requirements (i) to (v), for example:

a) the **Angoff Yes/No method** or the **standard Angoff method**
Either of these could be carried out in the initial series, probably using a two-stage process. The first stage could take place prior to the award (as in current GCE OTs) with the awarders sending in their responses for each item. At the award, standard item facility information could be provided, along with item facilities organised by total mark (as per Fearnley (2003) / MacCann and Stanley (2006) – although probably based on CTT, i.e. for each total mark the actual proportion correct per item)[31]. The grade boundary mark suggested by the awarders' responses could be compared with the SRB and the item facilities on those marks and in the range between considered in order to establish the final boundary mark.

In future series, the awarders could be provided with the SRB, overall item facilities, and the facilities per item in a five- or three-mark range centred on the SRB. Their consideration of these facilities (in relation, say, to how the awarders would have expected each item to perform) and the cumulative percentages on each mark would inform their final decision on the final recommended boundary.

b) the **Bookmark** approach or **item mapping**
A standard Bookmark or item mapping approach could be carried out in the initial series, for the former possibly using CTT rather than Rasch IRT to order the items in the item booklet, but in either case the booklet preferably being generated automatically. The first round of the standard Bookmark approach could potentially be carried out prior to the award, with the awarders sending in their bookmarks in advance. At the award, standard item facility information could be provided, along with item difficulties for each total mark achievable on the test (i.e. as per Fowles (2006), but potentially using CTT facilities). The grade boundary suggested by the awarders' bookmarks could be compared with the SRB and the 67% items (or items on an alternative agreed success criterion) for these two marks and the marks in between used to help establish the committee's recommendation for the final grade boundary.

In future series, the awarders could be provided with the SRB, overall item facilities and the actual item difficulties on marks in a five- or three-mark range centred on the SRB. Again the 67% items would be used to focus the discussion and establish the final recommended boundary.

Which of these options[32] is the correct answer? That notorious question allied to multiple-choice questions pertains and quite possibly in this case we will have to accept that there is no single correct answer.

Lesley Meyer
Senior Research Officer

December 2008

[31] In fact it would be possible to provide overall item facilities prior to the award, given that these data are usually available further in advance than summary data for written papers or coursework components.
[32] if any!

6. REFERENCES

Angoff, W. H. (1971). Scales, norms and equivalent scores. *In R. L. Thorndike (Ed.) Educational Measurement (2nd ed.) 508-600. Washington DC: American Council on Education.*

Baird, J. & Dhillon, D. (2005). Qualitative expert judgements on examination standards: valid but inexact. *AQA internal report, RPA_05_JB_RP_077.*

Balch, W. R. (1989). Item order affects performance in multiple-choice exams. *Teaching of Psychology,16(2), 75-77.*

Béguin, A., Kremers, E. & Alberts, R. (2008). National Examinations in the Netherlands: standard setting procedures and the effects of innovations. *Paper presented at the IAEA Conference, Cambridge, September 2008.*

Beretvas, N. S. (2004). Comparison of Bookmark difficulty locations under different item-response models. *Applied Psychological Measurement, 28(1), 25-47*.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56(1), 137-172.*

Buckendhahl, C. W., Smith, R. W., Impara, J. C. & Plake, B. S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement, 39(3), 253-263.*

Chinn, R. N. & Hertz, N. R. (2002). Alternative approaches to standard setting for licensing and certification examinations. *Applied Measurement in Education, 15(1), 1-14.*

Cizek, G. J. (1996). Setting passing scores. *Educational Measurement: Issues and Practice, 15(2), 20-31*.

Cizek, G. J. (Ed.) (2001). Setting Performance Standards: Concepts, methods and perspectives. *Mahwah, NJ: Lawrence Erlbaum Associates.*

Cizek, G., Bunch, M. B. & Koons, H. (2004). Setting performance standards: contemporary methods. *Educational Measurement: Issues and Practice, 23(4), 31-50.*

Cizek, G. J. & Bunch, M. B. (2007). Standard setting: a guide to establishing and evaluating performance standards on tests. *Sage Publications Inc.*

Clauser, B. E., Swanson, D. B. & Harik, P. (2002). Multivariate generalizability analysis of the impact of training and examinee performance information on judgements made in an Angoff-style standard-setting procedure. *Journal of Educational Measurement, 39(4), 269-290.*

Cross, L. H., Impara, J. C., Frary, R. B. & Jaegar, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement, 21, 113-129.*

De Gruijter, D. N. M. (1985). Compromise models for establishing examination standards. *Journal of Education Measurement, 22, 263-269*.

Ebel, R. L. (1972). Essentials of educational measurement. *Englewood Cliffs, NJ: Prentice-Hall.*

Fearnley, A. J. (2003). An investigation into the possible application of item response theory to provide feedback of information to awarders in the use of Angoff's method of standard setting in AQA OTQ components. *AQA internal report RC_240*.

Fowles, D. (2003). Standard Setting: a review of some recent approaches to setting boundary marks on the basis of examiner judgement. *AQA internal report, RC_207.*

Fowles, D. (2004). A trial of a bookmark approach to grading and comparisons with the Angoff method. *AQA internal report, RC_259.*

Fowles, D. (2005). A further trial of a bookmark approach to grading objective tests. *AQA internal report, RPA_05_DEF_RP_07.*

Green , D. R., Trimble, C. S. & Lewis, D. M. (2003). Interpreting the results of three different standard setting procedures. *Educational Measurement: Issues and Practice, 22(1), 22-32.*

Hambleton, R. K., Brennan R. L., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., Nelhaus, J., Reckase, M. D., Rindone, D., van der Linden, W. J. & Zwick, R. (2000). A response to "Setting reasonable and useful performance standards" in the National Academy of Sciences' Grading the Nation's Report Card. *Educational Measurement: Issues and Practice, 19(2), 5-14.*

Hambleton, R. K. & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8(1), 41-55.*

Huck, S. W. & Bowers, N. D. (1972). Item difficulty level and sequence effects in multiple-choice assessments tests. *Journal of Educational Measurement, 9(2), 105-111.*

Hurtz, G. M. & Hertz, N. R. (1999). How many raters should be used for establishing cutoff scores with the Angoff method: a generalizability theory study. *Educational and Psychological Measurement, 59, 885-897.*

Impara, J. C. & Plake, B. S. (1997). Standard setting: an alternative approach. *Journal of Educational Measurement, 34(4), 353-366.*

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

Jaegar, R. M. (1982). An iterative structured judgement process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4, 461-475.*

Karantonis, A. & Sireci, S. G. (2006). The Bookmark standard-setting method: a literature review. *Educational Measurement: Issues and Practice, 25(1), 4-12.*

Laffitte, R. G. Jr. (1984). Effects of item order in achievement test scores and students' perceptions of test difficulty. *Teaching of Psychology, 11(4), 212-213.*

MacCann, R. G. & Stanley, G. (2006). The use of Rasch modelling to improve standard setting. *Practical Assessment, Research & Evaluation, 11(2), 1-17.*

Mehrens, W. A. (1995). Methodological issues in standard setting for educational exams. *In Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments (p.p. 221-263). Washington DC: National Assessment Governing Board and National Center for Educational Statistics.*

Meyer, L. (2003a). AQA Standards Unit analyses for the GCE Economics awarding meeting, June 2003. *AQA internal report, RC_238.*

Meyer, L. (2003b). Repeat AQA Standards Unit analyses originally run for the GCE Economics awarding meeting, June 2003. *AQA internal report, RC_239.*

Mills, C. N. & Melican, G. J. (1988). Estimating and adjusting cutoff scores: features of selected methods. *Applied Measurement in Education, 1(3), 261-275.*

Mitzel, H. C., Lewis, D. M., Patz, R. J. & Green, D. R. (2001). The Bookmark procedure: psychological perspectives. *In G. J. Cizek (Ed.), Setting Performance Standards: Concepts, methods and perspectives. Mahwah, NJ: Lawrence Erlbaum Associates.*

National Academies of Sciences (2005). Measuring literacy: performance levels for adults, interim report. Appendix C: July 2004 Bookmark standard-setting session with the 1992 NALS data (pages 221-284).
Retrieved from http://www.nap.edu/openbook.php?record_id=11267&page=221.

Neely, D. L., Springston, F. J. & McCann, S. J. H. (1994). Does item order affect performance on multiple-choice exams? *Teaching of Psychology, 21(1), 44-45.*

Newman, D. L., Kundert, D. K., Lane, D. S. Jr. & Bull, K. S. (1988). Effect of varying item order on multiple-choice test scores: importance of statistical and cognitive difficulty. *Applied Measurement in Education, 1(1), 89-97.*

Perlini, A. H., Lind, D. L. & Zumbo, B. D. (1998). Context effects on examinations: the effects of time, item order and item difficulty. *Canadian Psychology, 39(4), 299-307.*

Schagen, I. & Bradshaw, J. (2003). Modelling item difficulty for bookmark standard setting. *Paper presented at the BERA annual conference, Herriott-Watt University, Edinburgh, 11-13 September 2003.*

Schultz, E. M., Lee, W. & Mullen, K. (2005). A domain-level approach to describing growth in achievement. *Journal of Educational Measurement, 42, 1-26.*

Schultz, E. M. & Mitzel, H. C. (April, 2005). The Mapmark standard setting method. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Stringer, N. (2007). Evaluation of the February 2007 alternative awarding procedure trials. *AQA internal report, RPA_07_NS_RP_039.*

Vos, P. & Kuiper, W. (2003). Predecessor items and performance level. *Studies in Educational Evaluation, 29, (191-206).*

Wang, N. (2003). Use of the Rasch IRT model in standard setting: an item-mapping method. *Journal of Educational Measurement, 40(3), 231-253.*

Yin, P. & Sconing, J. (2007). Estimating standard errors of cut scores for item mapping and mapmark procedures: a generalizability theory approach. *Educational and Psychological Measurement, 68(1), 25-41*.

Zieky, M. J. (2001). So much has changed: how the setting of cutscores has evolved since the 1980s. *In G. J. Cizek (Ed.), Setting Performance Standards: Concepts, methods and perspectives. Mahwah, NJ: Lawrence Erlbaum Associates.*

APPENDIX A:

The Bookmark method: Hypothetical output from round one of the Bookmark procedure (adapted from Cizek & Bunch, 2007, page 183)

	Basic		Advanced	
Participant	Page in OIB*	Ability (theta) at that cut point	Page in OIB*	Ability (theta) at that cut point
1	5	-0.334	46	1.627
2	8	0.082	46	1.627
3	8	0.082	47	1.650
4	6	-0.243	46	1.627
5	10	0.270	38	1.333
6	6	-0.243	39	1.340
7	9	0.193	40	1.489
8	6	-0.243	39	1.340
9	7	-0.176	40	1.489
10	8	0.082	40	1.489
11	7	-0.176	43	1.586
12	8	0.082	41	1.510
13	9	0.193	42	1.580
14	9	0.193	46	1.627
15	9	0.193	39	1.340
16	9	0.193	42	1.580
17	13	0.420	38	1.333
18	8	0.082	22	0.600
19	10	0.270	39	1.340
20	11	0.272	39	1.340
Mean cut		0.060		1.442
Median cut		0.082		1.489
Mean cut (on raw mark scale)		22.04		39.87
Median cut (on raw mark scale)		22.31		40.32

*OIB=Ordered Item Booklet

APPENDIX B:

The Rasch model

The basic premise underlying the Rasch IRT model is that an observed item response is governed by an unobservable candidate ability variable, θ, and the item difficulty. The probability of a candidate with an ability level of θ answering an item correctly can be modelled by:

$$P_{ij} = \frac{1}{1 + e^{-(\theta_i - \beta_j)}} \quad (1)$$

where θ_i is the ability of candidate *i*, and β_j is the difficulty of item *j*.

Rearranging equation (1), the estimated ability of candidate *i* on any item becomes a function of the probability of success on that item:

$$\theta_i = \beta_j - \ln\left(\frac{1 - P_{ij}}{P_{ij}}\right) \quad (2)$$

where *ln* is the natural logarithm (log to the base *e*).

i.e., $\theta_i = \beta_j + k$, where *k* is a constant.

Thus, using a response probability of 0.67 in (2)

$$\theta_i = \beta_j - \ln\left(\frac{1 - 0.67}{0.67}\right) = \beta_j + 0.693$$

Whereas, using a response probability of 0.50 in (2) yields

$$\theta_i = \beta_j - \ln\left(\frac{1 - 0.50}{0.50}\right) = \beta_j$$

(Source: MacCann & Stanley, 2006)

APPENDIX C (i):

A subset of items (from a 50 item test) arranged by item difficulty, as estimated from a one-parameter Rasch model

Item	Proportion correct	IRT item difficulty	Transformed item difficulty*
30	0.70	-0.861	37
23	0.70	-0.857	37
38	0.68	-0.767	37
14	0.67	-0.687	37
29	0.65	-0.537	38
45	0.63	-0.494	38
24	0.63	-0.468	38
8	0.61	-0.349	39
27	0.60	-0.338	39
10	0.60	-0.327	39
20	0.60	-0.324	39
1	0.59	-0.288	39
43	0.58	-0.244	39
28	0.57	-0.178	39
4	0.57	-0.136	39
35	0.56	-0.114	40
48	0.55	-0.064	40
37	0.54	0.008	40
40	0.53	0.043	40
46	0.47	0.323	41
44	0.46	0.384	42
3	0.43	0.582	42
7	0.40	0.758	43
15	0.39	0.772	43
39	0.36	0.897	44
13	0.34	1.007	44
33	0.34	1.029	44
9	0.32	1.141	45
17	0.26	1.561	46
26	0.25	1.630	47
12	0.25	1.702	47
49	0.22	1.803	47
2	0.23	1.821	47
18	0.19	2.066	48
11	0.19	2.231	49
6	0.16	2.280	49
16	0.17	2.442	50

* (IRT item difficulty x 4) + 40, rounded to the nearest integer
Source: MacCann & Stanley (2006), page 13

APPENDIX C (ii):

An item map, based on the transformed difficulties of items in (i)