# Thinking on the edge: the influence of discussion and statistical data on awarders' perceptions of borderline candidates in an Angoff awarding meeting

*Nadezda Novakovic*

## Abstract

*The Angoff method is a widely used procedure for setting pass scores in vocational examinations, in which the awarders estimate the performance of minimally competent candidates (MCCs) on each test item. Within the context of some UK vocational examinations, the procedure consists of two stages: after making the first round of estimates, awarders make final decisions after discussion and after receiving statistical data about candidate performance.*

This study investigated the relative effects of discussion and performance data on awarders' estimates relating to a UK vocational qualification. The results of the study showed that performance data had more influence on the awarders' estimates than discussion alone. However, neither discussion nor performance data had the power to drastically sway the awarders from their original decisions, nor did they significantly reduce the variability of individual judgments.

These data were compared to what the awarders reported in questionnaires and interviews. The comparison revealed that there was often a discrepancy between what the awarders said about the effect of discussion and statistics on their estimates and what one can conclude by looking at quantitative data alone.

## Background

During the past few decades, a surge in standardised testing within the educational contexts of North America and Australia has helped intensify the interest in various methods for setting pass-scores in criterion-referenced assessment. As a result, these methods have come under close scrutiny not only from the research community, but also from a wider community of stakeholders who have a vested interest in assuring that these are the most accurate and fair means of separating competent students from those whose performance has not yet reached the established standard in a given subject area.

One of the most widely used procedures for computing pass scores in both the vocational and general education settings is the Angoff method (Angoff 1971). It was originally devised as a standard-setting method, ideally to be applied only the first time the test is created. In this case, a panel of awarders (also referred to as judges, panellists or experts) with subject expertise are asked to individually estimate, for each test item, the percentage of *minimally competent* or *borderline* candidates (MCCs) who should answer that item correctly. A minimally competent candidate is defined as a candidate with sufficient skills to only just achieve a pass. These percentages are summed across items, and the result is an individual awarder's pass score for the test paper in question. The average of all individual awarders' scores represents the recommended pass mark for the test.

Within the context of UK OCR[1] examinations, the Angoff procedure is routinely used to set pass scores for some vocational qualifications that result in pass-fail decisions. In this context, the Angoff method is used not only for standard setting, but also for standard-maintaining purposes, to ensure year-on-year equivalency of pass scores. Consequently, the awarders are asked to make a prediction as to how many MCCs *would* get each question right, rather than a judgement as to how many *should* get a test item right. This implies a shift of focus from hypothetical students to the students that awarders are familiar with.

Furthermore, in the context of UK OCR qualifications, the awarders have the opportunity to make two rounds of estimates. They make the initial estimates individually, at home. Later on, at an awarding meeting, they discuss the perceived difficulty of test items. They also receive performance data in the form of item facility values, which represent the percentage of all candidates who answered each test item correctly. After discussion and presentation of performance data, the awarders make their final estimates about MCCs' performance. Both discussion and performance data are supposed to increase the reliability of the procedure by reducing the variability among individual estimates, and Busch & Jaeger (1990) provide some empirical evidence for this claim.

Among the main strengths of the Angoff method, Ricker (2006) lists its simplicity, the fact that it is easy to implement and explain to awarders and stakeholders, and that it uses simple statistics that are easy to calculate and understand.

However, the validity and reliability of the Angoff procedure have been questioned in recent literature. The main criticism is directed against the assumed high cognitive load of the awarders' task, who need to form a mental representation of a group of MCCs and estimate as accurately as possible the performance of such candidates on a test. Shepard (1995, cited in Plake & Impara 2001, p.88), argues that both of these task exceed the capacity of human raters, while the National Academy of Education[2] characterised Angoff and similar item-judgment methods as "fundamentally flawed because they require awarders to perform a nearly impossible cognitive task" (cited in Berk 1996, p.216).

Ricker (2006) warns against the possible dangers of conceptual drift, i.e. the inability of awarders to maintain the mental image of MCCs throughout the entire awarding activity. Laming's (2004) work in psychology provides ample evidence that humans indeed cannot maintain a stable frame of reference over longer periods of time.

Although Hambleton et al. (2000) argue that there is no psychological evidence to suggest that task of conceptualising MCCs is cognitively taxing, a number of studies suggest that awarders do indeed encounter many difficulties when making item performance estimates. Boursicot & Roberts (2006) found that experts in their study generally disagreed on the definition of borderline candidates and had trouble translating the concept of MCCs into item performance estimates. Hayes (2001), Impara & Plake (1998) and Sizmur (1997) all found that awarders' estimates are on the whole inaccurate and do not reflect well the performance of low achieving students. Hayes (2001) concludes that asking awarders to estimate probabilities is "a wholly unreasonable request", and that using pre-test data coupled with expert judgment is a better way of arriving at pass scores.

The reliability of the Angoff method has also been questioned: item performance estimates may be affected by individual awarders' own ideas about the competencies of MCCs, which may reduce the reliability of the procedure. In Boursicot & Roberts' (2006) study, for example, different panels of awarders set significantly different pass marks for the same tests. Giraud, Impara & Plake (1995) found that teachers have varied perceptions about the characteristics of MCCs which result in different pass scores.

Sometimes, especially in the US context, there have been attempts to overcome the problem of high variability by employing a large number of awarders, so that any large discrepancies in their estimates may be cancelled out in the final

calculation of the pass score. However, the literature is not clear on what the most appropriate number of awarders in an Angoff meeting is: Norcini & Shea (1997) recommend as few as five, while Cizek (1996) recommends using as many awarders as possible.

In the context of UK OCR vocational awards, employing a large number of awarders is usually not an option; discussion and performance data are therefore relied upon to reduce the variability of judgements. However, during discussion, awarders may feel pressure to conform to the opinion of the entire panel. This phenomenon is well documented in social psychology research (Asch 1951, Cartwright & Zander 1960, Sherif 1935) and Fitzpatrick (1984) collated evidence from a number of studies that this tendency is indeed present in the context of standard-setting and standard-maintaining procedures. Murphy *et al*. (1995) provided clear evidence of group conformity during discussions which are part of the UK awarding meetings for general education qualifications.

Ricker (2006) points to the potential danger of presenting awarders with performance data, if those data are obtained from an unrepresentative sample of candidates. Furthermore, performance data refer to the entire candidature for the given qualification, while awarders are asked to estimate the performance of minimally competent rather than all candidates.

Additionally, an increase in reliability which may be achieved by the introduction of performance data or discussion does not always equate with an increase in quality. According to Ricker (2006), awarders are expected to have different views about the desired performance standards, and the loss of individual opinions would be detrimental to the very essence of the judgmental process.

In short, the validity of the Angoff procedure rests on a very fine balance between increasing the agreement between awarders, while at the same time preserving the variety of individual opinions about candidate performance. While discussion and performance data may help with increasing inter-awarder reliability, they may also open the way for many biases to creep into the decision-making processes: awarders may feel pressure to conform to the opinions held by the group, and their final decisions may be unduly influenced by statistical information.

## Study aims

The study was conducted in the context of a UK OCR vocational qualification where the Angoff method was regularly used for standard maintaining purposes. The aim of the study was to investigate the relative effects of discussion and performance data on: (1) the awarders' perceptions of MCCs, (2) their expectations of how MCCs might perform on the test, (3) their consistency in basing their decisions on the performance of MCCs only, and not on the

performance of all or average candidates, (4) their level of confidence and (5) their ability to rank-order items in terms of their relative difficulty. In other words, the study is an attempt to link quantitative changes in the awarders' item performance estimates to any possible concomitant changes in their cognitive processes.

# Method

## *Design*

A group of seven awarders made item performance estimates for two tests of comparable difficulty. The first round of judgements for both tests was made individually, at home. Later on, the awarders attended two awarding meetings, one for each test. At the first meeting, the awarders discussed the perceived difficulty of each test item in turn. Following the discussion, the awarders made the final round of item performance estimates. The second meeting took place one hour after the first meeting; the awarders took part in a discussion, but they were also given the performance data before making the final round of estimates. The second meeting resembled as closely as possible the usual OCR Angoff awarding meetings for Vocational Qualifications. The fact that the awarders received performance data at only one of the meetings allowed us to tease apart the effect of discussion and performance data on their final decisions.

## *The awarding meetings*

The awarding meetings were chaired by an experienced OCR Chairperson. At the start of the first meeting the Chairperson introduced the Angoff procedure and the concept of a minimally competent candidate, whom he described as a student who would pass the test on a good day, but fail on a bad day. The awarders were told not to make estimates on whether MCCs *should* or *ought to* know the question, but on whether they *would* get the question right. They were also encouraged to think about students they had taught. This is a usual recommendation at the OCR Angoff awarding meetings, and while it may help reduce the cognitive difficulty of the awarders' task, it may increase the variability of their judgements.

In order to reduce the potential influence of more vocal awarders on the decisions of the rest of the panel, the awarders were asked not to mention during the discussion the exact estimate values they had given to the test items.

The awarders first voiced their opinions about the test paper in general, after which they discussed each item in turn. After each item was discussed, the awarders had the chance to change their original estimates, although there was no requirement for them to do so.

At the start of the second meeting, the Chairperson first explained the performance data that the awarders would get at the meeting, which included

the discrimination and facility indices for each item. It was made clear that the item facility values did not reflect the performance of MCCs, but the performance of the entire group of candidates who took Test 2. The Chairperson emphasised that there was no reason for the panel to agree with the item facility values, but he did mention that these were a good indicator of the relative difficulty of the test items.

After the introduction, the second meeting followed the same format as the first meeting.

## Tests

The tests used in the study were constructed from items used in a unit from the OCR Certificate in Teaching Exercise and Fitness Level 2 (Unit 1 – Demonstrate Knowledge of Anatomy and Physiology). These items were drawn from an item bank, and their Rasch difficulty values had already been established. This had several advantages. Firstly, it allowed the construction of two tests of comparable difficulty. Secondly, the pass mark could be established by statistical means, using the information on how students performed on these items in the past. The pass mark for both tests was set at 18.

Test 1, containing 27 items, was completed by 105 students, and Test 2, containing 28 items, was completed by 117 students from centres offering Teaching and Exercise qualification. The tests were completed as part of another experimental study (Johnson 2007), i.e. these were not 'real' tests and and were used only within an experimental context. The students who completed these tests were aware that their results would be used only for research purposes. Students completed Test 1 after completing Test 2.

It is important to note that the selection of students in Johnson's (2007) study can best be described as opportunistic. All the students came from eight centres that were willing to let their pupils take part in the study. The implications of this type of selection are revisited in the discussion section.

The awarders in our study were not informed whether the tests they had been given were live tests or not. However, they were aware that the awarding meetings they were attending were purely experimental and that their decisions would be used only for research purposes.

## Awarders

The awarding panel consisted of three female and four male awarders, all experts in the field of Teaching Exercise and Fitness. The recruitment process focussed on the small pool of experts who had already taken part in Angoff awarding meetings for this qualification. Since only five of these experts were available to take part in the study, we additionally recruited two more awarders

who, although without any previous experience of the Angoff procedure, fulfilled all the necessary criteria (in terms of professional experience and expertise) to take part in an Angoff awarding meeting.

## Questionnaire and interviews

A questionnaire and semi-structured telephone interviews were used to collect information about the awarders' perceptions of MCCs, the strategies they used to make item estimates, as well as their own views about the influence of discussion and performance data. The qualitative data collection was unobtrusive and retrospective: the questionnaires were completed after the first round of estimates at home, and the semi-structured telephone interviews were conducted a day after the awarding meetings. The awarders were thus able to complete their task uninterrupted.

Both the questionnaire and the interviews comprised a combination of direct questions, open-ended questions and rating scales. Skorupski & Hambleton (2005) used a questionnaire as a means of investigating the cognitive processes of a group of Angoff raters by asking them to share their thoughts at five different points during the procedure (in the context of an English Language test). Among other things, their questionnaire elicited the awarders' thoughts about the influence of discussion and feedback on their item ratings, confidence levels and their understanding of the student performance levels. Some of these questions have been adapted to the purposes of our study to elicit the awarders' views on the influence of discussion and performance data in the context of a UK OCR vocational award.

The questionnaire asked only about the first phase of the study (work at home), while interview questions referred to all the stages of the procedure. The results section focuses in detail on these questions, which provided information about the awarders' perceptions of the difficulty of their tasks ( i.e. how difficult they found it to form a mental image of MCCs and to estimate their performance), the strategies they used to perform these tasks, as well as their level of confidence in their estimates. The awarders also provided their own views on the extent to which they felt their perceptions of MCCs and performance estimates were influenced by discussion, performance data and the pressure to agree with the rest of the panel.

Additionally, the awarders were asked which group of candidates they had in mind when making the estimates: whether they thought about MCCs, average candidates or all candidates. To our knowledge, the consistency with which the awarders base their decisions on the performance of MCCs has not been addressed in the literature, although Ricker (2006) expressed the concern that Angoff awarding decisions may suffer from conceptual drift.

Additional questions elicited information about the awarders' level of

concentration and fatigue at both meetings to ensure that the results of the study were not influenced by these external factors.

All questions are reported upon in the results section of this paper except the questionnaire items that elicited information about the awarders' personal information, experience and background. The remainder of the interview and questionnaire items are addressed and discussed in the results section of the paper), and it is therefore not considered necessary to reproduce the survey instrument here.

## Minimally competent candidates

In order to compare the awarders' estimates to the actual performance of MCCs, we had to identify the group of MCCs from all the candidates who took the tests. Remember that the awarders' estimates are supposed to reflect the percentage of minimally competent candidates rather than the percentage of all candidates who would answer test items correctly.

MCCs were identified as those candidates whose score fell 1 SEM[3] (Standard Error of Measurement) above and 1 SEM below the pass score calculated by using the item bank data. This is a method similar to the one used in Goodwin (1999) and Plake & Impara (2001). However, in these studies the authors used a circular approach: the accuracy of awarders' estimates was measured against the performance of a group of candidates identified as MCCs by using the pass mark set by the very same awarders whose accuracy was being measured. In our study, this problem was avoided, as the pass mark obtained from item bank data provided a more objective means of identifying MCCs.

The first column of Table 1 shows the pass marks for both tests calculated by using item bank data. The second and third columns show the mean score achieved by all candidates and the group of candidates we identified as MCCs respectively. Figures in brackets represent the percentage of the total possible score.

*Table 1 The average performance of all candidates and MCCs on Tests 1 and 2* [4]

| | | All candidates | | MCCs | |
|---|---|---|---|---|---|
| | Pass mark | Mean mark | N | Mean mark | N |
| Test 1 | 18 (67%) | 17.60 (65%) | 105 | 17.87 (66%) | 38 |
| Test 2 | 18 (66%) | 16.04 (57%) | 117 | 17.57 (63%) | 46 |

On the whole, the performance of MCCs was better than the average performance of all candidates. Also, all candidates performed better on Test

1 than on Test 2. Johnson (2007) ascribed this to the practice effect, since the candidates completed Test 1 after Test 2. However, four members of the awarding panel (Awarders 1, 4, 6 and 7) voiced their opinion that Test 2 was harder than the usual tests administered for this qualification.

## Findings

The awarders' estimates were compared to the actual item facility values for the group of MCCs by using the following measures: mean actual difference (MD), mean absolute difference (MAD), and Spearman rank order correlation coefficient (Spearman rho).

Actual differences were calculated by subtracting the observed item facility values from the awarders' estimates. Positive values indicate that, on average, an awarder expected MCCs to perform better than they actually did, while negative values indicate that their expectations were lower than the actual MCC performance.

Absolute differences were calculated in the same way as actual differences, but were all assigned positive values. Absolute differences provide a clear indication of the size of the difference between the awarders' estimates and the actual item facility values.

Spearman rank-order correlation coefficient was used as a measure of the extent to which the awarders' ranking of items matched the rank-ordering of items in terms of their actual facility values (i.e. how easy or difficult the items actually were for the group of borderline candidates).

The following sections report on the awarders' estimates during each stage of the standard setting procedure, as well as on their own thoughts and experiences during each of these stages. The results of statistical analyses have already been reported in Novakovic (2008), although this is the first time that the quantitative and qualitative data have been presented together.

## Phase 1: Working at home

The awarders made the first round of estimates individually, at home. All the data in this section refer to this specific stage of the awarding procedure.

In the questionnaire, the awarders were asked how difficult they found it to conceptualise MCCs. The awarders chose responses from a six-item rating scale ranging from "very difficult" to "very easy". Only one awarder, Awarder 1, said she found it somewhat difficult, while the other six, including the two inexperienced awarders, said they found it easy or very easy.

When asked in the interviews what specific strategies they used to imagine a group of MCCs, all the awarders mentioned using their own students as

a reference point. Thus, the awarders used this strategy even before being advised to do so at the awarding meetings. Impara & Plake (1997) speculated that thinking about familiar students is less cognitively demanding than thinking about hypothetical candidates; Ferdous *et al.* (2006) and Skorupski & Hambleton (2005) found that this was indeed a common strategy used by awarders in their studies. However, basing judgments on a group of familiar students may increase the variability among individual estimates, as illustrated by the following quote:

> "[Our students are] very, very motivated, so the standard generally is very high. So I find it difficult to see the minimally competent, so I go across […] all the courses that we've done and sort of visualise people from there." (Awarder 7)

The comment reveals how the standard or the ability of the specific students the awarders teach may colour their perception of what MCCs in general would be able to do.

While the awarders generally reported finding it easy to imagine a group of MCCs, the interview data show that they had more problems in estimating their performance. Choosing a response from a five-item rating scale, ranging from "very difficult" to "very easy", only Awarder 2 said he found the task of estimating MCC performance easy, while six awarders admitted that they found it somewhat difficult. This is a finding similar to the one in Boursicot & Roberts (2006), who found that awarders have less difficulty in forming a concept of MCCs than in translating this concept into a numerical score.

Despite the difficulties, the awarders generally reported high levels of confidence in their first round of estimates (Figure 1). They rated their confidence levels by using a five-item rating scale ranging from "very confident" to "very unconfident". [5]

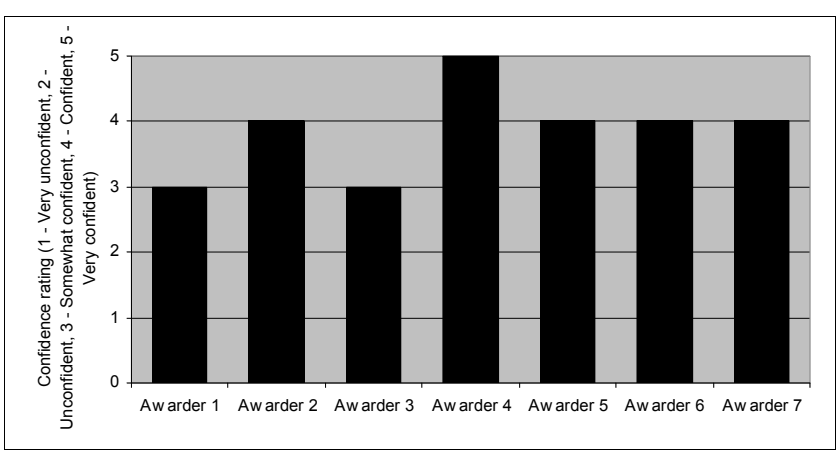


*Figure 1 Self-reported confidence levels after making first round of estimates*

In the interviews, the awarders reported on strategies they used in order to make item estimates at home (Table 2). Note that the numbers in Table 2 and the following tables do not always add up to seven: sometimes, some of the awarders may have given more than one answer, and some awarders may not have answered all the questions.

*Table 2 Strategies used for making item performance estimates*

| Response | N |
|---|---|
| Thinking about how one's own students would answer the question | 4 |
| Thinking about what one finds difficult to teach and/or understand oneself | 1 |
| Thinking about what is being taught and revised with students on a regular basis | 1 |
| Concentrating on the quality of distractors | 2 |
| Going for the "gut reaction" | 1 |

Most of the awarders tried to imagine themselves in the position of familiar students, although they did not specify whether they thought about low achievers or average students in their class. However, the awarders' responses identified some potentially less legitimate strategies for predicting the MCC performance. For example, thinking about what one finds difficult to teach may distort the judgement of what candidates themselves find easy or hard to do. Going for the "gut reaction" is also a potentially problematic strategy; it implies a hastily made decision which may have been arrived at without giving due consideration to the competencies of a very specific group of candidates.

In the interviews, the awarders were presented with three statements asking which group of candidates they thought about while making estimates: MCCs, average candidates or all candidates. The awarders chose a response from a five-item scale ranging from "strongly agree" to "strongly disagree" for each of these statements. While all awarders agreed or strongly agreed that they thought about MCCs, Awarder 4 mentioned thinking about all candidates, and Awarders 3 and 6 reported thinking about average candidates as well. Basing decisions on candidates other than MCCs is a threat to the validity of the Angoff method. A comment from Awarder 6 illustrates this; the quote also shows how basing estimates on gut reactions may not be the most appropriate decision-making strategy.

> "I suppose you do think of the type of thing that a lot of students find difficult […] so that that could be an average to a certain extent, couldn't it? Where you're going for a gut reaction, you're possibly going

more for an average kind of thing. It's when you start analysing particularly difficult questions, that's when you come back to your minimally competent, yes. So perhaps there is a bit of average going on in there as well." (Awarder 6)

Table 3 contains mean item performance estimates for each awarder. Mean item performance estimates represent the percentage of the total possible mark, while figures in brackets (last row) represent this percentage translated into the recommended pass mark. Awarders 3 and 5 are the inexperienced awarders.

*Table 3 Individual awarders' mean item performance estimates for Tests 1 and 2*

| | Test 1<br>Mean estimates | Test 2<br>Mean estimates |
|---|---|---|
| Awarder 1 | 73.70 | 70.36 |
| Awarder 2 | 71.11 | 69.29 |
| Awarder 3 | 80.19 | 81.43 |
| Awarder 4 | 82.59 | 64.64 |
| Awarder 5 | 72.59 | 65.54 |
| Awarder 6 | 75.74 | 65.71 |
| Awarder 7 | 81.85 | 78.57 |
| Mean ( all awarders) | 76.82 (21) | 70.79 (20) |

What is immediately obvious from the figures is the variability in awarders' judgments, which is not unexpected if one takes into consideration that awarders based their decisions on the performance of different groups of students using a variety of strategies.

However, the mean estimates also show that the awarders generally agreed in their expectations that MCCs would perform better than the group of candidates we identified as borderline actually did. This is especially true of Test 1, where all the awarders' mean estimates were higher than the mean mark achieved by MCCs (17.87 or 66 % of the total mark). On Test 2, the expectations were still high, although some awarders' mean estimates were closer to the mean mark achieved by MCCs (17.57 or 63% of the total mark). This could be ascribed to the fact that some awarders felt Test 2 to be harder (than Test 1) and consequently lowered their expectations for this test.

Table 4 shows the correlation between the awarders' estimates and the actual item facility values for both tests: the awarders were better in predicting the relative difficulty of test items on Test 2 than on Test 1.

*Table 4 Spearman rank-order correlations between estimated and actual MCC item facility values*

| | Round 1 estimates |
|---|---|
| Test 1 actual item facility values (MCCs) | .234 |
| Test 2 actual item facility values (MCCs) | .601** |

**P <0.01

# Phase 2: awarding meetings

## *Angoff awarding meeting 1: Influence of discussion*

During the first meeting of the day, the awarders took part in discussion, after which they had the chance to change their initial estimates for Test 1 items.

The average number of changes to the initial estimates was only 5.14 (remember that Test 1 consisted of 27 test items). Awarders 1 and 7 made most changes (10 and 9 respectively), while the inexperienced Awarder 5 made no change to any of her initial estimates. Furthermore, there was no change to the pass mark from the first round of estimates – it remained at 21.

The analysis of mean actual (MD) and mean absolute (MAD) differences between the estimated and actual MCC item facility values indicates that the discussion did not have much influence on the awarders' final estimates. Figure 2 shows the mean actual differences for Test 1 on both rounds.

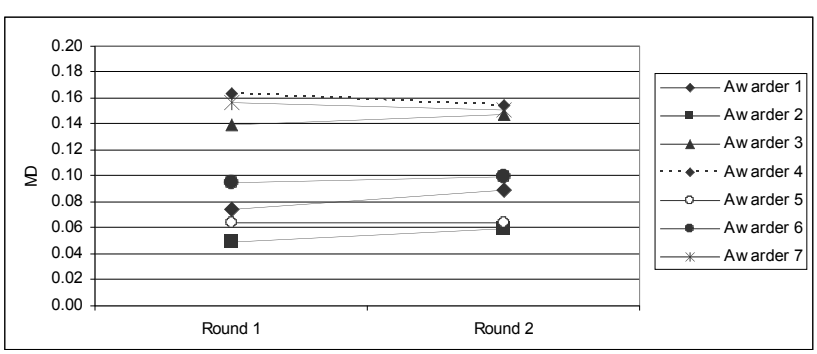


*Figure 2 The MDs between estimated and observed item facility values on two rounds of estimates for Test 1*

A split-plot ANOVA revealed a significant main effect of awarder ($F(6) = 12.87$, $p < 0.001$), but no significant interaction between round and awarder ($F(6) = 0.13$, $p = 0.99$) nor any significant main effect of round ($F(1) = 0.12$, $p = 0.73$). This means that overall the examiners made similar estimates on the two rounds.

Figure 3 shows the mean absolute differences for both rounds of estimates for Test 1.

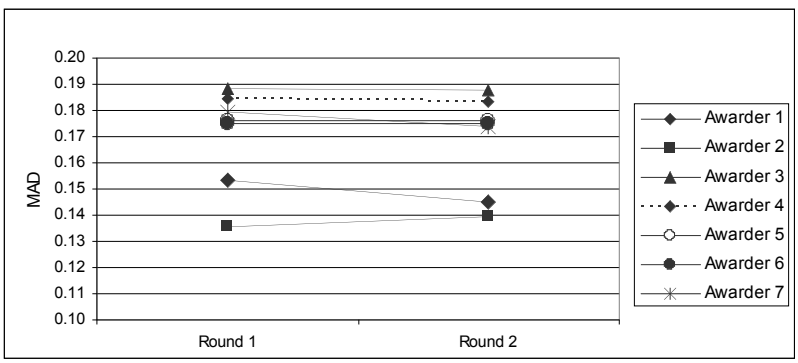


*Figure 3 The MADs between estimated and observed item facility values on two rounds of estimates for Test 1*

A split-plot ANOVA revealed a significant main effect of awarder ($F(6) = 2.83$, $p = 0.01$). There was no main effect of round ($F(1) = 2.26$, $p = 0.13$) nor any significant interaction between round and awarder ($F(6) = 0.85$, $p = 0.53$). These results indicate that the awarders, while differing among themselves in the size of the MAD, made similar judgments on both rounds.

Furthermore, as shown by figures in Table 5, the discussion did not seem to improve the correlation between the estimated and actual MCC item facility values.

*Table 5 Spearman rank-order correlations between estimated and actual MCC item facility values for Test 1*

| | Test 1 actual item facility values (MCCs) |
|---|---|
| Round 1 estimates | .234 |
| Round 2 estimates | .187 |

Despite the results of quantitative analysis suggesting otherwise, the awarders felt that discussion did indeed have some influence on both their perception of MCCs as well as on their estimates.

Except for Awarder 2, who felt that discussion had no influence at all, all the other awarders confirmed that discussion influenced their perception of MCCs; their comments are summarised in Table 6.

*Table 6 Influence of the discussion on awarders' perception of MCCs*

| *Response* | *N* |
|---|---|
| Reinforcing the concept of MCC/ helping focus on MCCs | 5 |
| Confirming the original decisions | 1 |
| Giving different aspects/ideas | 4 |

Generally, the awarders felt the influence of discussion to be positive and beneficial; it served as an opportunity to hear, for the first time during the procedure, different opinions about the abilities of MCCs against which they could compare their own views. This in turn helped them focus their attention on the minimally competent, as illustrated by the comments below.

> *"It reinforces the minimally competent when you're with other people who have different experience of different learners and things as well."* (Awarder 6)

> *"[…] Let's say I thought that the students would find something really easy and everybody else thought it was a really difficult question, […] and you have to think 'well yes, if that's what the majority of people think, that's probably a good point'. So it does influence you, seeing how other people rate it."* (Awarder 6)

The last comment also hints at a tendency of some of the awarders to agree with the majority of the panel. This is however not reflected in the quantitative data, which do not show, for any of the awarders, any significant changes to the original decisions after the discussion.

The awarders' answers as to which group of candidates they thought about while making final Test 1 estimates also indicate that the awarders were more focussed on the MCCs after the discussion. All awarders agreed or strongly agreed that they thought about MCCs, and only Awarder 4 reported thinking

about all candidates as well. A comment by Awarder 6 illustrates how the discussion oriented her decision making from the average towards the borderline candidates:

> *"I think possibly, thinking less of average students after the discussion. So when I was at home I was probably going a little bit more on average student but then when having the discussion and seeing other people's input... No I was thinking more of minimally competent." (Awarder 6)*

When asked whether they agreed that discussion influenced their final estimates for Test 1, the awarders chose responses ranging from "strongly agree" to "strongly disagree". Six awarders agreed with the statement, although three awarders agreed with it only somewhat. Awarder 2 disagreed. However, the results of the quantitative analysis showed that while the awarders made some changes to their original estimates, these did not reach statistical significance.

When asked directly about it, six of the awarders said they never or rarely felt pressure to agree with other awarders, except for Awarder 1 who said she felt this pressure at times. The awarders also reported being confident or very confident in their estimates (Figure 4).

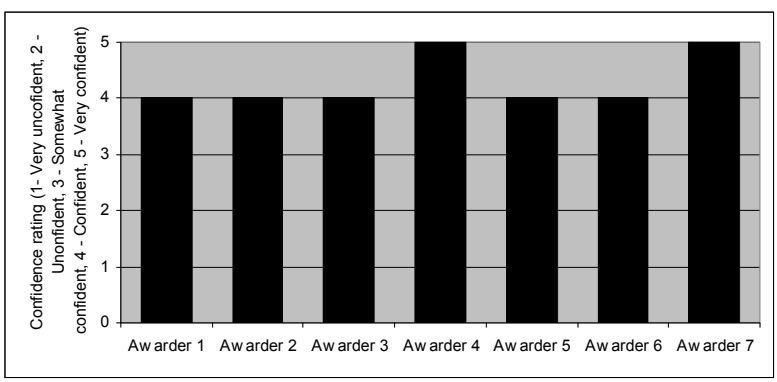


*Figure 4 Self-reported confidence levels after making second round of estimates for Test 1*

## *Angoff awarding meeting 2: Influence of performance data*

At the second meeting, the awarders took part in a discussion and received the information on how the students performed on each test item.

Table 7 compares the number of changes that the awarders made to the initial estimates at each meeting: several awarders made more changes at the second meeting, and the average number of changes increased from 5.14 to 11.29. However, despite the changes, the recommended pass mark for Test 2 changed only by one mark from the first round of estimates (from 20 to 19).

*Table 7 Number of changes to the initial item performance estimates*

| | No. of changes for Test 1 | No. of changes for Test 2 |
|---|---|---|
| Awarder 1 | 9 | 18 |
| Awarder 2 | 3 | 9 |
| Awarder 3 | 4 | 22 |
| Awarder 4 | 6 | 8 |
| Awarder 5 | 0 | 1 |
| Awarder 6 | 4 | 16 |
| Awarder 7 | 10 | 5 |
| Mean for all awarders | 5.14 | 11.29 |

Figure 5 shows the mean actual differences for Test 2 at both rounds.

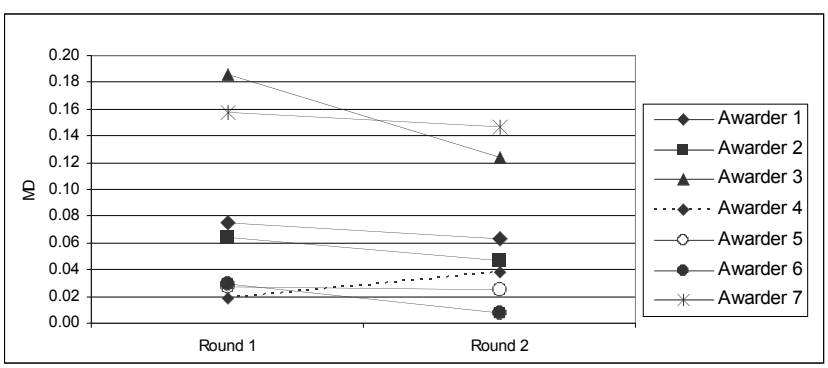


*Figure 5 The MDs between estimated and observed item facility values on two rounds of estimates for Test 2*

A split-plot ANOVA revealed a significant main effect of awarder ($F(6) = 18.79$, $p < 0.001$), but there was no main effect of round ($F(1) = 2.26$, $p = 0.13$) nor any significant interaction between round and awarder ($F(6) = 0.85$, $p = 0.53$). Although the change was not statistically significant, Figure 6 shows that it was rather pronounced for the inexperienced Awarder 3, who made most changes to his initial estimates.

Figure 6 shows the mean absolute differences for Test 2 on both rounds.

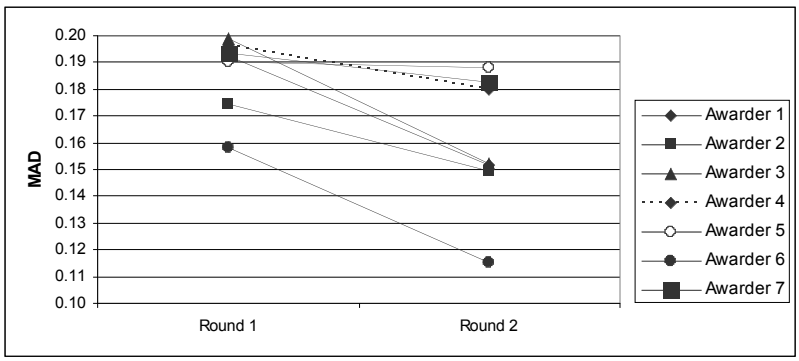


*Figure 6 The MADs between estimated and observed item facility values on two rounds of estimates for Test 2*

The ANOVA results for Test 2 revealed a significant main effect of awarder ($F(6) = 2.29$, $p = 0.036$). The interaction between round and awarder was not significant ($F(6) = 0.13$, $p = 1$), but there was a significant main effect of round ($F(1) = 7.76$, $p = 0.005$): the mean difference between rounds was 0.026, which is a large effect size ($d = 1.3$). This indicates a statistically significant change in the size of the MAD between two rounds for Test 2, which was not revealed for Test 1.

Furthermore, the correlations between the awarders' estimates and the actual item facility values for Test 2 became slightly stronger after the second round of estimates (Table 8). This contrasts the situation from the first meeting.

*Table 8 Spearman rank-order correlations between estimated and actual MCC item facility values for Test 2*

| | Test 2 actual item facility values (MCCs) |
|---|---|
| Round 1 estimates | .601** |
| Round 2 estimates | .793** |

**P <0.01

When asked what they themselves thought about the influence of discussion and performance data at the second meeting, five out of seven awarders agreed that discussion influenced their item performance estimates. Awarder 2 said that discussion had no influence on his final decisions, while Awarder 5 was undecided.

When asked about the influence of performance data, six awarders agreed or strongly agreed that these data influenced their final decisions, with the exception of the inexperienced Awarder 5, who disagreed. This is confirmed by the quantitative data which show that this awarder made the least number of changes to her initial decisions. Table 9 summarises the awarders' comments about the influence of performance data.

*Table 9 The influence of performance data on awarders' final decisions*

| *Response* | *N* |
|---|---|
| Surprise effect | 2 |
| Undermining confidence | 2 |
| Making you rethink | 1 |
| Helpful | 1 |

While one of the awarders found the performance data helpful, the others' comments were less favourable. Unlike the discussion at the first meeting of the day, which mostly seemed to confirm their original decisions, the performance data made the awarders question the estimates they had made at home. This explains partly the large number of changes to their original estimates. The comments below illustrate this; the last comment reveals how one of the awarders even doubted the veracity of the presented statistics.

> *"I think where we didn't have the students' results, […] I didn't make as many changes […] because I was still confident with what I said. But my confidence perhaps is undermined a little bit when you see the students' response, you think 'oh I didn't get that right', you know. "* (Awarder 6)

> *"I think when we got the statistics, […] I just couldn't see how any of those statistics were accurate. Not from […] the experience I've had with my students. So I wondered whether they'd been put in there as a kind of red herring…so that did influence me".* (Awarder 3)

When asked which group of candidates they thought of while making estimates, all the awarders agreed or strongly agreed that they were thinking about MCCs. Awarder 4 reported thinking about all candidates, and Awarders 1 and 6 reported thinking about average candidates as well. Interestingly, Awarder 6 is the awarders who reported that discussion at the first meeting helped her focus on MCCs. A comment below reveals her feeling that performance data may sway the decision-making process towards average, rather than borderline candidates:

> *"I think it's this average performance of the learners, you know the statistics, it's a big influence, that. And the chairperson did keep reminding us that that was the average person and that what you've got to focus on is the minimally competent […] I think that's important because you could lose track and you could get caught up in […] the fact that it is the average student and get that confused with the minimally competent."* (Awarder 6)

When asked about peer pressure, Awarders 1, 2, 5 and 7 said they were never pressured to agree with other awarders, while Awarder 4 felt this pressure sometimes. Awarders 3 and 6 felt this pressure often; these are the same awarders who reported that their confidence was undermined by the presentation of performance data..

Figure 7 shows the awarders' self-reported confidence levels during the second meeting of the day. The inexperienced awarder 3 reported feeling unconfident in his estimates. This is the same awarder who often felt the pressure to agree with the other members of the awarding panel, and who also made the most changes to his original estimates (22 out of 27).

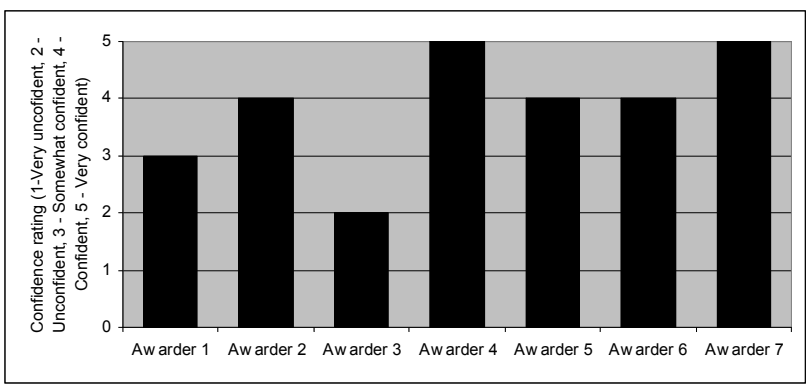


*Figure 7 Self-reported confidence levels after making second round of estimates for Test 2*

# Discussion

In this study, we sought to address the respective influences of discussion and performance data on the decisions made during Angoff awarding meetings in the context of a UK vocational qualification, where the Angoff method is standardly used for standard-maintaining purposes. In the light of various criticisms that have been put forward against the Angoff method, our aim was to investigate whether discussion and performance data have the power to increase the validity and reliability of the Angoff procedure by focussing on the following: (1) the awarders' perceptions of MCCs, (2) their expectations of MCC performance, (3) their consistency in making decisions based on MCCs,

(4) their confidence levels and (5) their ability to rank-order items in terms of their relative difficulty. A combination of quantitative and qualitative data was used to address these issues.

## Findings relating to phase 1 (work at home)

The findings relating to the first phase of the study indicated that awarders reported little difficulty in conceptualising MCCs, and that their strategy of choice was to think about familiar students. While they were able to identify the low achievers among the familiar students, the awarders found the task of estimating MCC performance more challenging; this confirms the findings by Boursicot & Roberts (2006) and Impara & Plake (1998). Despite difficulties in estimating item facilities, the awarders felt relatively confident while working at home.

The awarders' comments revealed that the standard and ability of specific groups of students they are working with could colour and potentially skew their perception of how MCCs in general would perform on a test. This may at least partly explain the results of statistical analysis, which revealed statistically significant differences among the individual awarders' estimates. These differences were not unexpected however; after all, the purpose of discussion and performance data is to smooth out these differences and bring awarders' individual estimates in line with each other. Furthermore, despite individual differences, all the awarders expected MCCs to perform better on both tests than the group of candidates we identified as MCCs actually did.

An unwelcome finding was that at least some of the awarders used somewhat inappropriate strategies for making item performance estimates when working alone, such as, for example, 'going for the gut reaction' or 'basing estimates on average candidates'. This was unexpected, especially as the awarders were specifically instructed to think about MCCs, and as majority of the participants in the study were experienced Angoff awarders. This findings emphasises the need for the training of awarders before the awarding meetings take place.

## Findings relating to the influence of discussion (first meeting)

The main finding form the qualitative strand of the analysis suggests that the awarders felt that discussion influenced both their understanding of the abilities of MCCs as well as their item performance estimates. The awarders appreciated the opportunity to hear other views on how the MCCs would perform on certain test items. Most importantly, the discussion oriented the decision-making process, for at least some of the awarders, from the average towards the MCC candidates, thus potentially increasing the validity of the procedure.

The discussion seemed to provide a non-threatening environment in which to share views and reassess one's decision: this is supported by the fact that the awarders' confidence levels remained high after the discussion, and that they felt little if any pressure to agree with the rest of the panel.

However, despite the self-reported effect of discussion on their decisions, the statistical analysis did not reveal any significant change from the first round of estimates, either in the direction or the magnitude of differences between the awarders' estimates and the actual MCC performance. The correlation between the awarders' estimates and the actual MCC item facility values also showed no improvement after the discussion.

Thus, the discussion did not seem to bring the awarders' judgments into line, a finding contrary to Busch & Jaeger (1990). This is potentially problematic for small qualifications with smaller awarding panels, where the discussion is expected to 'iron out' differences among awardees and bring their estimates toward a more unified picture of MCC performance. In short, the findings suggest that discussion on its own may not be the best possible way of increasing the reliability and accuracy of awarders' judgments.

## Findings relating to the influence of discussion and performance data together

The awarders' comments suggest that the performance data mostly challenged the awarders' original decisions; some awarders expressed surprise at and even incredulity over the statistics provided, and two awarders felt that their confidence was undermined by these data. In general, the discussion coupled with the presentation of performance data seemed to be more threatening for some awarders than discussion alone. This is supported by the fact that several awarders felt pressure to agree with the rest of the panel at the second meeting.

The comments also revealed that, at least for some of the awarders, the performance data made them think more about average than borderline candidates. This is not altogether surprising, and suggests yet another potential area of difficulty facing the awarders: making use of performance data referring to the entire candidature in order to make estimates about the performance of borderline candidates only. This finding gives support to the concerns about possible conceptual drift expressed by Ricker (2006).

However, the presentation of performance data did improve the accuracy of awarders' estimates. The changes the awarders made to the original estimates resulted in a statistically significant decrease in the magnitude of differences between the estimated and actual MCC item facility values. However, the awarders who made the most changes were the ones whose self-reported

confidence levels were the lowest at the second meeting; they were also most inclined to change their estimates even when not agreeing with the statistics. This reveals that the presentation of performance data has the potential to create a pressure environment especially for those without enough experience and/or confidence.

The performance data, however, did not significantly influence the direction of differences between the awarders' estimates and the actual MCC item facility values. The pass mark changed very little from the first round of estimates and on the whole the awarders still overestimated the performance of MCCs. This indicates that most of the awarders did not blindly follow the statistics, but used them more to 'fine-tune' their original judgments in view of the new evidence.

## Limitations

The experimental design of the study was such that only one awarding panel judged both tests, with a risk that the study could be suffering from order effects. Having two panels judging both tests in a different order would be a definite improvement to the present design. Although we had hoped to involve two groups of awarders, we were unable to recruit enough participants for this study.

Although the tests used in the study were supposed to be of the same difficulty, the students performed better on one of the tests. Having two groups of students completing the tests in different order would have provided a better indication of whether the better performance on one of the tests was due to the practice effect or whether it could be ascribed to the inherent difficulty of the tests.

Considering that the selection of students in the original study (Johnson 2007) was opportunistic and depended on the willingness of centres to take part in research, it is acknowledged that the performance data obtained may not be fully representative. Furthermore, the students who completed the tests were aware that their results would be used only for research purposes; it is possible that their performance would have differed if they were taking a live test.

Also, the study focussed on only one qualification, and it is conceivable that the results of the study might have been different if a different qualification was used in the experiment or if the awarding panel was different. However, we believe that the participants in the study reflect well the experience and expertise of other awarders who take part in OCR Angoff awarding meetings for various vocational qualifications.

## Conclusions

The results of study suggest that discussion and performance data may be useful additions to the Angoff standard-setting procedure, if both are present at the awarding meeting. The opportunity to discuss the perceived difficulty of test items may help the awarders focus on MCCs and provides them with different views about the performance of such candidates. Furthermore, it is viewed positively by the awarders themselves, and seems to increase their confidence. However, on its own, the discussion does not seem to have the power to bring the awarders' estimates into line or improve either their accuracy or the rank-ordering of the test items.

On the other hand, the presentation of statistical data led to a decrease in the magnitude of differences between estimated and actual MCC differences and had a positive effect on the awarders' ability to rank-order items by their relative difficulty. However, the performance data did not have the same influence on all the awarders, and those whose confidence was low seemed to be most affected by the presentation of these statistics. However, even after the presentation of performance data, the awarders still overestimated the performance of MCCs.

Furthermore, the study has highlighted the difficulties one faces in trying to determine the validity of (any aspect of) the Angoff method. The main problem is a sort of Catch 22 situation that researchers find themselves in: the awarders are asked to set the standard for a specific qualification, but then one sets out to test their accuracy without the existence of any real external reference point. In this study, we used item bank data to test the accuracy of the awarders' decisions, but these are not always available and the Angoff method is usually used when one lacks data of this type. It is therefore imperative that the Angoff procedure is subjected to rigorous comparisons with other standard-setting and standard-maintaining methods. Such continuous investigations are necessary to ensure that the methods used are the most reliable, valid and fair means of assessing candidates' competence in any given subject area.

## Notes

1. OCR (Oxford, Cambridge and RSA) is one of the major providers of qualifications in the UK.

2. National Academy of Education is an independent US honorary society which, in its own words, "advances the highest quality education research and its use in policy formation and practice".

3. The Standard Error of Measurement estimates how repeated measures of a person on the same instrument tend to be distributed around their "true" score – the score that they would obtain if a test were completely error-free.

4. Table 1 and Figures 2, 3, 5 and 6 are reproduced from Novakovic (2008) with the kind permission of the editor of *Research Matters: A Cambridge Assessment Publication.*

5. The awarders provided information about their confidence levels during the first round of estimates both in the questionnaires and in the interviews. Since there was no discrepancy between the answers given on two occasions, the paper presents the information obtained from the interviews.

**Nadezda Novakovic** is a Research Officer with the Research Division of Cambridge Assessment in the UK. Email: Novakovic.N@cambridgeesol.org

## References

Angoff, W 1971, *Scales, norms and equivalent scores,* American Council on Education, Washington, DC.

Asch, S E 1951, 'Effects of group pressure upon the modification and distortion of judgments', in *Groups, Leadership and Men,* ed H Guetzkow, Carnegie Press, Pittsburgh.

Berk, R 1996, 'Standard setting: the next generation (where few psychometricians have gone before!)', *Applied Measurement in Education*, vol 9, pp.215-235.

Boursicot, K & Roberts, T 2006, 'Setting standards in a professional higher education course: Defining the concept of the minimally competent student in performance based assessment at the level of graduation from medical school', *Higher Education Quarterly*, vol 60, pp.74-90.

Busch, J & Jaeger, R 1990, 'Influence of type of judge, normative information, and discussion on standards recommended for the National Teachers Examinations', *Journal of Educational Measurement*, vol 27, no.2, pp.145-163.

Cartwright, D & Zander, A 1960, *Group dynamics* (2nd ed), Row, Peterson and Company, Evanston IL.

Cizek, G 1996, 'Setting passing scores', *Educational Measurement: Issues and Practice*, vol 15, no.2, pp.20-31.

Ferdous, A, Nering, M & Plake, B 2006, 'Factors that influence judges' decisions in an Angoff standard setting study', Paper presented at the 2006 American Educational Research Association Annual Meeting, San Francisco, CA.

Fitzpatrick, A 1984, 'Social influences in standard-setting: The effect of group interaction on individuals' judgement', Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Giraud, G, Impara, J & Plake, B 2005, 'Teachers' conceptions of the target examinee in Angoff standard setting', *Applied Measurement in Education*, vol 18, no 3, pp223-232.

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

Laming, D 2004, *Human judgment: The eye of the beholder,* Thomson, London.

Hambleton, R K, Brennan, R L, Brown, W, Dodd, Forsyth, R A, Mehrens, W A, Nellhaus, J, Reckase, M, 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,* vol 19, no.2, pp.5-14.

Hayes, M 2001, 'The role of the Angoff procedures in the level setting processes for the end of Key Stage tests in mathematics', AGRAQ paper, Qualifications and Curriculum Authority internal report.

Impara J & Plake, B 1997, 'Standard setting: an alternative approach', *Journal of Educational Measurement*, vol 34, no.4, pp.353-366.

Impara, J & Plake, B 1998, 'Teachers' ability to estimate item difficulty: a test of the assumptions in the Angoff standard setting method', *Journal of Educational Measurement*, vol 35, no.1, pp.69-81.

Johnson, M 2007, 'Does the anticipation of a grade motivate vocational test takers?', *Research in Post Compulsory Education,* vol 12, no.2, pp.159-179.

Murphy, R, Burke, P, Cotton, T, Hancock, J, Partington, J, Robinson, C, Tolley, H, Wilmut, J, Gower, R 1995, 'The dynamics of GCSE awarding', Report of a project conducted for the School Curriculum and Assessment Authority', School of Education, University of Nottingham.

Norcini, J & Shea, J 1997, 'The credibility and comparability of standards', *Applied Measurement in Education*, vol 10, no.1, pp.39-59.

Novakovic, N 2008, 'The influence of performance data on awarders' estimates in an Angoff awarding meetings', *Research Matters: A Cambridge Assessment Publication,* vol 5, pp.15-19.

Plake, B & Impara, J 2001, 'Ability of panelists to estimate item performance for a target group of candidates: an issue in judgmental standard setting', *Educational Assessment*, vol 7, no.2, pp.87-97.

Ricker, K 2006, 'Setting cut-scores: a critical review of the Angoff and modified Angoff methods', *The Alberta Journal of Educational Research*, vol 52, no.1, pp.53-64.

Shepard, L A 1995, 'Implications for standard setting of the National Academy of Education evaluation of National Assessment of Educational Progress achievement levels', *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments*, U.S. Government Printing Office, Washington DC.

Sherif, M 1935, 'A study of some social factors in perception', *Archives of Psychology,* vol 22, no.187.

Sizmur, S 1997, 'Look back in Angoff: a cautionary tale', *British Educational Research Journal*, vol 23, no.1, pp.3-13.

Skorupski, W & Hambleton, R 2005, 'What are panelists thinking when they participate in standard-setting studies?', *Applied Measurement in Education*, vol 18, no.3, pp.233-356.