Dunning, D., Heath C. & Suls, J.M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, **5**, 3, 69–106.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, **34**, 906–911.

Greatorex, J. & Suto, W.M.I. (2005). *What goes through a marker's mind? Gaining theoretical insights into the A-level and GCSE marking process*. A report of a discussion group at Association for Educational Assessment – Europe, Dublin, November 2005.

Griffin, D. & Tversky, A. (2002). The weighing of evidence and the determinants of confidence. In: T. Gilovich, D. Griffin. & D. Kahneman (Eds.) *Heuristics and biases: The psychology of intuitive judgement*. Cambridge: Cambridge University Press, 230–249.

Koch, Adina (2001). Training in metacognition and comprehension of physics texts. *Science Education*, **85**, 6, 758–768.

Maki, R. H. (1998). Test predictions over text material. In: D.J. Hacker, J. Dunlosky & A.C. Graesser (Eds.) *Metacognition in educational theory and practice*. London: Lawrence Erlbaum Associates, 117–144.

Murphy, R., Burke P., Cotton, T. *et al.* (1995). *The dynamics of GCSE awarding. Report of a project conducted for the School Curriculum and Assessment Authority*. Nottingham: School of Education, University of Nottingham.

Qualifications and Curriculum Authority (2006). *GCSE, GCE, VCE, GNVQ and AEA Code of Practice 2005/6*. London: Qualifications and Curriculum Authority.

Schraw, G. (1998). Promoting General Metacognitive Awareness. *Instructional Science*, **26** 113–25.

Sherif, C., Sherif, M. & Nebergall, R. (1965). *Attitude and attitude change: The social judgement-involvement approach*. Philadelphia: Saunders.

Suto, W.M.I. & Greatorex, J. (*in press*). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policies and Practice*.

Suto, W.M.I. & Nádas, R. (2007a). The 'Marking Expertise' projects: Empirical investigations of some popular assumptions. *Research Matters: A Cambridge Assessment Publication*, **4**, 2–5.

Suto, W.M.I. & Nádas, R. (2007b). *What makes some GCSE examination questions harder to mark than others? An exploration of question features related to marking accuracy*. A paper presented at the British Educational Research Association Annual Conference, London, 2007.

Suto, W.M.I. & Nádas, R. (*in press*). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education*.

Weinstein, N.D (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, **39**, 806–820.

ASSESSMENT JUDGEMENTS

The influence of performance data on awarders' estimates in Angoff awarding meetings

Nadežda Novaković Research Division

Background

A variety of standard-setting methods are used in criterion-referenced assessment[1] to decide upon pass scores which separate competent from not yet competent examinees. During the past few decades, these methods have come under close scrutiny not only from the research and academic community, but also from a wider community of stakeholders who have a vested interest in assuring that these methods are the most accurate and fair means of determining performance standards.

The Angoff method (Angoff, 1971) is one of the most widely used procedures for computing cut scores in both the vocational and general education settings. In the Angoff standard setting procedure, a panel of judges with subject expertise are asked to individually estimate, for each test item, the percentage of *minimally competent* or *borderline* candidates (MCCs)[2] who would be able to answer that item correctly.

Within the context of some OCR multiple-choice vocational examinations, judges have the opportunity to make two rounds of estimates. The awarders make the initial estimates individually, at home. Later on, they attend an awarding meeting, at which they take part in a discussion about the perceived difficulty of test items. Furthermore, the awarders receive performance data in the form of item facility values, which represent the percentage of all candidates who answered each test item correctly. Both discussion and performance data are supposed to increase the reliability of the procedure and help judges make more accurate estimates about the performance of MCCs (Plake and Impara, 2001).

After discussion and presentation of performance data, the awarders make their final estimates as to what percentage of MCCs would answer each test item correctly. These percentages are summed across items, and the result is an individual judge's pass score for the test paper in question. The average of individual judges' scores represents the recommended pass mark for the test.

The Angoff method is popular because it is flexible, easy to implement and explain to judges and stakeholders, and it uses simple statistics that are easy to calculate and understand (Berk, 1986; Goodwin, 1999; Ricker, 2006).

However, the validity and reliability of the Angoff procedure have been questioned in recent literature. The main criticism is directed against the high cognitive load of the task facing the awarders, who need to form a mental representation of a hypothetical group of MCCs, maintain this image throughout the entire standard setting activity, and estimate as accurately as possible how a group of such candidates would perform on

[1] In criterion-referenced assessment, a candidate's performance is judged against an externally set standard.

[2] A minimally competent or a borderline candidate is a candidate with sufficient skills to only just achieve a pass.

a test (Berk, 1996; Boursicot and Roberts, 2006; Glass, 1978; Impara and Plake, 1997; Plake and Impara, 2001).

Some of the criticism has also been directed against the potential undesirable effects of discussion and performance data. During the discussion, awarders may feel pressure to conform to the opinion of the group (Fitzpatrick, 1984, cited in Busch and Jaeger, 1990), while performance data from a small unrepresentative sample of candidates may introduce flaws into the procedure (Ricker, 2006). Furthermore, performance data refer to the entire candidature for the given qualification, while judges are asked to estimate the performance of *minimally competent* rather than *all* candidates. Additionally, some researchers have warned that reliability may be artificially introduced by performance data or discussion by eliminating the variability of individual judgements, whereby the resulting standard may 'no longer reflect judges' true perceptions about the examinee performance' (McGinty, 2005; Ricker, 2006).

Aim

The aim of the study was to investigate the relative effect of discussion and performance data on: (1) the awarders' expectations on how MCCs might perform on a test, (2) the magnitude of change in the awarders' estimates between sessions and (3) the awarders' rank-ordering of items in terms of their relative difficulty.

Design

A group of seven awarders made item facility estimates for two tests of comparable difficulty. They made the first round of judgements for both tests individually, at home. At a later stage, the awarders attended two awarding meetings, one for each test. The meetings took place on the same day. At the first meeting, the awarders voiced their opinions about the quality of Test 1, after which they discussed the perceived difficulty of each test item in turn. Following the discussion, the awarders made the final round of item facility 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 item facility estimates.

The awarding meetings

The awarding meetings were chaired by an experienced Chairperson, who co-ordinated the procedure and facilitated the discussion in the way it is usually done at the OCR Angoff awarding meetings for Vocational Qualifications.

At the start of the first meeting the Chairperson introduced the Angoff procedure and the concept of a minimally competent candidate. He described an MCC as a student who would pass the test on a good day, but fail on a bad day. He also mentioned various ways which could help awarders conceptualise MCCs, for example, thinking about students they had taught. In other words, the awarders were directly encouraged to make estimates about the performance of candidates familiar to them. This is a usual recommendation at the OCR Angoff awarding meetings, and while it helps reduce the cognitive difficulty of the awarders' task, it may result in an increase in the variability of awarders' judgements. The awarders were also told not to make estimates on whether MCCs *should* or *ought* to know the question, but on whether they *would* get the question right.

The awarders were also asked not to mention during the discussion the exact estimate values they had given to the items, although they could say whether they had given a low or a high estimate. This recommendation was given to help reduce the potential influence of more vocal awarders on the decisions of the rest of the panel.

The awarders first voiced their opinions about the test paper in general and its relative difficulty and quality, 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 explained the statistical data that the awarders would get at the meeting, which included the discrimination and facility indices for each item. The awarders were made aware 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 the fact that there was no reason for the panel to make their item facility estimates agree with the actual item facility values, but he did mention that the latter were a good indicator of which question was easier or harder compared to other questions in the test.

After the introductory part, the second meeting followed the same format as the first meeting.

Tests

The tests used in the study were two multiple-choice tests constructed from the items used in Unit 1 of 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 IRT (Rasch) difficulty values had already been established. This had several advantages. First, 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 Exercise and Fitness qualification. The tests were completed as part of another experimental study (Johnson, *in press*), that is, these were not 'real' tests and student performance data were used only for research purposes. Students completed Test 1 after completing Test 2.

Awarders

The awarding panel consisted of three female and four male awarders. These were all experts in the field of Teaching Exercise and Fitness. Two awarders had no experience with the Angoff procedure, while the remaining five had already taken part in an Angoff awarding meeting.

Minimally competent candidates

In order to measure how the awarders' estimates compared to the actual performance of MCCs, we had to identify this group of candidates 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 (standard error of measurement) above and 1 SEM below the pass score[3] established by using the item bank data. This is a method similar to the one used in Goodwin (1996) and Plake and Impara (2001).

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

Table 1 : The average performance of all candidates and MCCs for Tests 1 and 2

	Pass mark	*All candidates*		*MCCs*	
		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 all candidates was better on Test 1 than Test 2. Johnson (*in press*) ascribed this to the practice effect, since the candidates completed Test 1 after having completed Test 2. However, it is worth noting that four members of the awarding panel voiced their opinion that Test 2 was harder than the usual tests administered for this qualification.

Key findings

Frequency of changes

The awarders made more changes to their original estimates if presented with statistical information about candidate performance than if they only took part in the discussion about the quality and perceived difficulty of the test items. The average number of changes between two rounds of estimates for Test 1 was 5.14 (ranging from 0 to 10 changes per awarder). For Test 2, however, the average number of changes was 11.29, with individual awarders making between 1 and 22 changes.

Rank-ordering of test items

The Spearman rank-order correlation coefficient was used to compare the awarders' estimates to the actual item facility values for the group of candidates identified as minimally competent. This showed how successful the awarders were in predicting which test items the MCCs would find harder and which ones they would find easier to answer. The correlation between the initial estimates for Test 1 and the actual item facility values was weak and non-significant (0.23), and it became weaker after the awarding meeting, at which the awarders took part only in discussion (0.19) On the other hand, the correlation between the initial estimates for Test 2 and the actual item facility values was significant and moderate (0.60), and it became stronger after the second meeting (0.79), when the awarders were presented with performance data. These findings are similar to the ones in Busch and Jaeger (1990), where correlations between the actual item facilities and mean item recommendations increased from one session to the other, after the awarders were presented with statistical information on students' performance.

Awarders' expectations

Table 2 shows the recommended pass marks, calculated by averaging the individual awarders' mean item facility estimates after each round of estimates for both Tests 1 and 2. The figures in brackets represent the percentage of the total possible mark.

Table 2 : The awarding panel's recommended pass marks for Tests 1 and 2 on two rounds of estimates

	Mean mark (all candidates)	*Mean mark (MCCs)*	*Recommended pass mark (Round 1)*	*Recommended pass mark (Round 2)*
Test 1	17.60 (65%)	17.87 (66%)	21 (77%)	21 (77%)
Test 2	16.04 (57%)	17.57 (63%)	20 (71%)	19 (69%)

Table 2 shows that, on average, the awarders' expectations were higher than the actual performance of the group of candidates we identified as minimally competent, as well as the entire group of students who took the test. This applies to both rounds of estimates.

Figures 1 and 2 show the mean actual difference (MD) between the awarders' estimates and the actual item facility values for the group of MCCs on both rounds, for Tests 1 and 2 respectively. The MDs were calculated by subtracting the observed item facility value from the awarder's estimated value. Positive values indicate that, on average, an awarder has mostly overestimated, while negative values indicate that the awarder has mostly underestimated the performance of MCCs. The graphs confirm that the awarders generally expected MCCs to perform better on both tests than they actually did, as indicated by the positive values of the individual MDs.

In order to see whether there was a statistically significant difference between the individual awarders' estimates on each round, an ANOVA was carried out on the data using the following model: 'Actual difference = round + item + awarder + round*item + awarder*round' (the asterisk sign, *, indicates an interaction between two variables).

The ANOVA results for Test 1 revealed that there was a significant main effect of item ($F(26) = 46.30$, $p < 0.001$), and a significant main effect of awarder ($F(6) = 12.87$, $p < 0.001$). There was no significant main effect of round ($F(1) = 0.12$, $p = 0.73$); the mean difference between two rounds was 0.003, which is a small effect size ($d = 0.06$), suggesting that overall the examiners made similar estimates on the two rounds. Furthermore, the analysis yielded no significant interaction between round and awarder ($F(6) = 0.13$, $p = 0.99$).

[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.

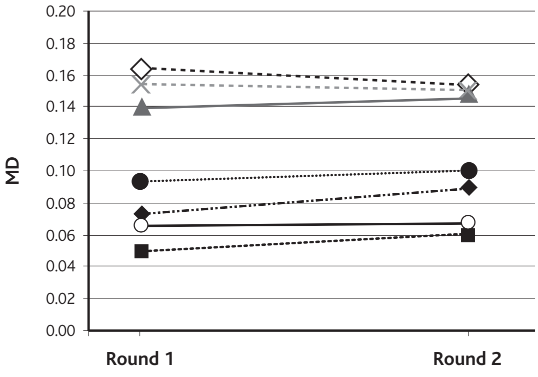

Figure 1: The MD between estimated and actual item facility values on two rounds of estimates for Test 1

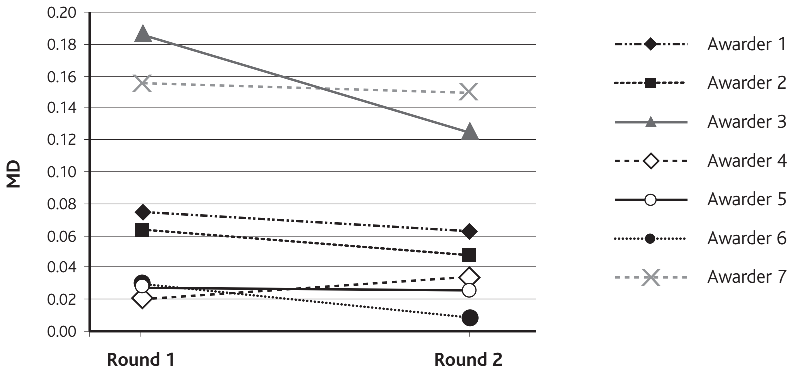

Figure 2: The MD between estimated and actual item facility values on two rounds of estimates for Test 2

On Test 2, the ANOVA revealed a significant main effect of item (F(27) = 44.75, p < 0.001) and a significant main effect of awarder (F(6) =18.79, p < 0.001), indicating that there was a statistically significant difference between individual awarders' MDs. There was no main effect of round (F(1) = 2.26, p = 0.13); the mean difference between the rounds was 0.015, which is a small effect size (d = 0.25). There was no significant interaction between round and awarder (F(6) = 0.85, p = 0.53).

Figures 3 and 4 show the mean absolute differences (MAD) between the awarders' estimates and the actual item facility value for the group of candidates we identified as minimally competent. Absolute differences were also calculated by subtracting the observed item facility values from the awarder's estimated item facility values. However, all differences were 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.

For Test 1, the results of an ANOVA with MAD as a dependent variable revealed a significant main effect of item (F(26) = 23.65, p < 0.001) and a significant main effect of awarder (F(6) = 2.83, p = 0.01). The main effect of round was not significant (F (1) = 0.04, p = 0.84); the mean difference between rounds was 0.002, which is a small effect size (d = 0.11). There was no interaction between round and awarder (F (6) = 0.03, p = 1.00).

The ANOVA results for Test 2 revealed a significant main effect of item (F(27) = 17.30, p < 0.001), and a significant main effect of awarder (F(6) = 2.29, p = 0.04). There was also 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). There was no significant interaction between round and awarder (F (6)= 0.13, p =1). These results revealed that overall there was a statistically significant change in the size of the MAD between two rounds, although there was no statistically significant difference in the way this changed for different awarders.

Conclusions and implications

The results of the present study support the current OCR practice that awarders at Angoff meetings should be presented with statistical data about candidates' performance.

If the awarders took part only in discussion about the perceived difficulty of test items, the number of changes the awarders made to their initial estimates was relatively small, and there was no change to the pass mark calculated using the initial estimates. Also, there was no statistically significant change from one round to the other, either in the direction or the magnitude of differences between the awarders' estimates and the actual performance of MCCs. Furthermore, the correlation between the awarders' estimates and the actual item facility values for MCCs became weaker after the discussion.

On the other hand, the combination of discussion and performance data had more effect on the awarders' estimates. After being presented with performance data, the awarders made, on average, twice as many changes to their original estimates than when they took part in discussion only. These changes resulted in a statistically significant decrease in the magnitude of differences between the awarders' estimates and the actual item facility values for the group of MCCs.

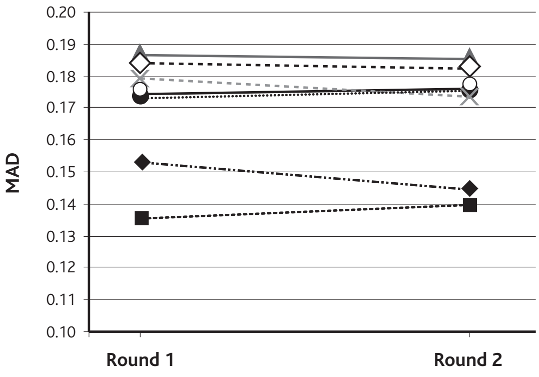

Figure 3 : The MAD between estimated and actual item facility values on two rounds of estimates for Test 1

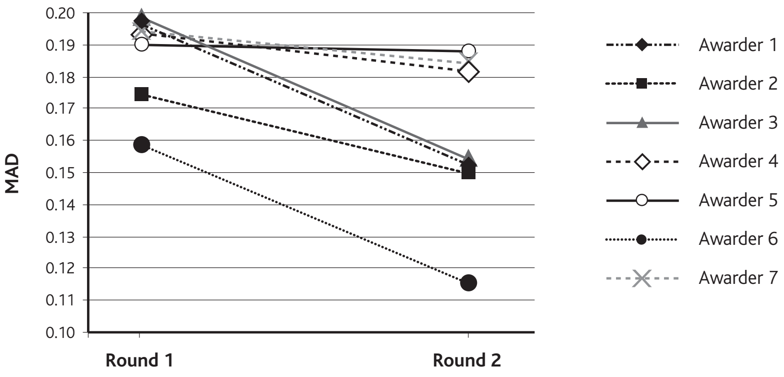

Figure 4 : The MAD between estimated and actual item facility values on two rounds of estimates for Test 2

Furthermore, after being provided with statistical data, the correlation between the awarders' estimates and the actual item facility values became stronger, indicating that the combination of statistical data and discussion helped the awarders judge the relative difficulty of the test items better. However, the provision of performance data had no impact on the direction of differences between the awarders' estimates and the actual item facility estimates.

An important aspect of these findings is that the changes made to the original estimates are observable mostly at the item level. In other words, while the awarders made changes to their item facility estimates, the actual change to the recommended pass mark was rather small (it decreased by only one mark). Furthermore, even after the second round, the recommended pass mark remained three marks higher than the average mark achieved by the total group of candidates. This indicates that the awarders were not swayed in their judgement by statistical data referring to the performance of the total group of candidates who took the test.

Another important finding is that the provision of statistical data does not seem to have affected the variability of awarders' judgements, a concern expressed by some researchers (McGinty, 2005; Ricker, 2006). Generally, if there was a statistically significant difference between the awarders, this difference was observable both before and after the provision of statistical data. In other words, the differences between the awarders were present even after they made changes to their original estimates, indicating that they still maintained their own views about how borderline students would perform on the test, regardless of the actual statistical data they received.

Although the study has provided important and useful findings, there were limitations which must be taken into account when considering its results. The influence of statistical data was tested on only one group of judges who made estimates about test items from a particular examination. However, we do believe that the members of the awarding panel chosen for the study reflect well the experience and expertise of other awarders who take part in the OCR Angoff awarding meetings for various vocational qualifications.

The experimental design of the study was such that only one awarding panel judged both tests, which means there is a risk that the design could be suffering from order effects. Having two awarding 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 unfortunately not able to recruit enough participants for this study. Furthermore, the fact that the awarders took part in discussion at both meetings could mean that the discussion they had at the first meeting influenced their judgements at the second meeting as well.

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.

It is important to note that the study focused only on some of the aspects of the Angoff method, without attempting to address the broader issues of the validity and reliability of the entire Angoff awarding procedure. These issues could be addressed by rigorous comparison of the Angoff method to other standard setting methods, such as the Bookmark method, for example. Such continuous investigations are necessary to ensure that methods used for setting pass scores are the most reliable, valid, fair and hence the most appropriate to be used both in the context of OCR vocational qualifications, as well as in the context of any standard-based examinations.

References

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

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

Berk, R. (1996). Standard setting: the next generation (where few psychometricians have gone before!). *Applied Measurement in Education*, **9**, 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*, **60**, 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*, **27**, 2, 145–163.

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.

Glass, G. (1978). Standards and Criteria. *Journal of Educational Measurement*, **15**, 237–61.

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

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

Jaeger, R. & Busch, J. (1984). *The effects of a Delphi modification of the Angoff-Jaeger standard-setting procedure on standards recommended for the National Teacher Examinations*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Johnson, M (*in press*). Does the anticipation of a grade motivate vocational test takers? *Research in Post Compulsory Education*.

McGinty, D. (2005). Illuminating the "black box" of standard setting; an exploratory qualitative study. *Applied Measurement in Education*, **18**, 3, 269–287.

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*, **7**, 2, 87–97.

Ricker, K. (2006). Setting Cut-Scores: A Critical Review of the Angoff and Modified Angoff Methods. *The Alberta Journal of Educational Research*, **52**, 1, 53–64.