Nádas, R. & Suto, W.M.I. (2007). *An exploration of self-confidence and insight into marking accuracy among GCSE maths and physics markers*. Paper presented at the annual conferences of the British Educational Research Association, 5–8 September, London; and the International Association for Educational Assessment, 17–21 September, Baku, Azerbaijan.

Pula, J. J. & Huot, B. (1993). A model of background influences on holistic raters. In: M. M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: theoretical and empirical foundations*. Cresskill, NJ: Hampton.

Sanderson, P. J. (2001). *Language and differentiation in Examining at A Level*. PhD Thesis. University of Leeds, Leeds.

Suto, W.M.I & Greatorex, J. (2006). A cognitive psychological exploration of the GCSE marking process. *Research Matters: A Cambridge Assessment Publication* **2**, 7–11. This article summarises key findings from Suto, W.M.I & Greatorex, J. (*in press*).

Suto, W.M.I & Greatorex, J. (*in press*, a). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*. Findings from this paper were also presented at the annual conference of the British Educational Research Association, September 2005, University of Glamorgan.

Suto, W.M.I & Greatorex, J. (*in press*, b). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policies and Practice*. Findings from this paper were also presented at the annual conference of the Association for Educational Assessment -Europe, November 2005, Dublin, Ireland.

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. Paper also presented at the annual conference of the International Association for Educational Assessment, 17–21 September, Baku, Azerbaijan.

Suto, W.M.I. & Nádas, R. (2007b). *What makes some GCSE examination questions harder to mark accurately than others? An exploration of question features related to accuracy*. Paper presented at the annual conference of the British Educational Research Association, 5–8 September, London.

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*.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In: L.Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts*. Norwood, N.J.: Ablex Publishing Corporation.

PSYCHOLOGY OF ASSESSMENT

An exploration of self-confidence and insight into marking accuracy among GCSE maths and physics markers

Rita Nádas and Dr Irenka Suto Research Division

Background

Introduction

A considerable volume of literature in education and occupational research investigates issues in self-confidence and insight, ranging from college students' post-diction self-assessment (e.g. Maki, 1998; Koch, 2001) to work-related self-assessment (Dunning, Heath and Suls, 2004). However, GCSE markers' perceptions of their marking performance and their metacognition have not, to our knowledge, been examined.

Exploring markers' perceptions is important for several reasons. First, if markers' estimates of their own performance prove to be accurate, then this information could be used by Awarding Bodies in standardisation procedures[1] to identify and discuss examination questions that markers have difficulties with. If, however, markers' insight proves to be unreliable and unrelated to their actual marking accuracy, then their feedback on 'problem areas' could be misleading: for example, when conducting standardisation procedures, Principal Examiners might find themselves focussing on the 'wrong' questions. Secondly, investigating whether self-confidence and insight change or become more accurate with more marking practice or more feedback could inform the marker training practices of Awarding Bodies. This may thereby enhance marking accuracy: there is evidence that improvement of one's self-assessment or insight into performance results in enhanced test performance (Koch, 2001; Dunning, Johnson, Ehrlinger and Kruger, 2003).

In this article we present the aims and findings of research which explored GCSE markers' perception of their own marking performance, namely, marking accuracy. Markers' levels of self-confidence and insight and possible changes in these measures over the course of the marking process were investigated. The term 'self-confidence' here denotes markers' post-marking estimates of how accurately they thought they had marked a sample of questions; 'insight' refers to the relationship between markers' actual marking accuracy and estimated accuracy, indicating how precise their estimates were.

Theories of insight and self-confidence

Insight into performance has been widely researched from various angles; and it has generally been found that people tend to have incorrect estimations of their own performance. For example, Dunning *et al*. (2003) found that when asked to predict their mastery on an examination, students in the bottom quartile greatly overestimated their actual performance. They also found that the better performing students were able to predict their raw scores with more accuracy, with top performers actually slightly underestimating their scores.

Several theories have been proposed to explain the phenomenon of poor insight. The nature of self-confidence has been examined by cognitive psychologists, who have adopted the 'self-serving bias' theory. Researchers have found that biases are used by participants in research

1 For regulations on standardisation procedures, see Qualifications and Curriculum Authority, 2006

situations in order to enhance or maintain positive self-views; for example, the *above average* effect (Dunning, Meyerowitz and Holzberg, 2002), or the *optimistic bias/unrealistic optimism* effect (for example, Armor and Taylor, 2002) have been described. Generally, it was found that people tend to have 'overinflated views of their skills that cannot be justified by their objective performance' (Dunning *et al*., 2003).

In some studies, participants were asked to estimate the probability of positive or negative life events that might happen to them (Weinstein, 1980); or to predict their own performance in an imagined or future situation, or *before* completing a task (for example, Griffin and Tversky, 2002). However, participants' actual performances were often not observed in these studies, or feedback was not provided. Thus, studies on self-serving self-assessments have not explored *change* in one's self-confidence after receiving feedback on actual performance. In the few studies in which participants' estimates were compared with their actual performances, results were mixed: while some found that performance estimates and actual performance did not correlate significantly (Griffin and Tversky, 2002), significant, positive and substantial correlations were found by others (e.g., when subjects made correct time estimates for a given task in the study of Buehler *et al*., 1994).

The self-serving bias theory alone cannot explain all findings. It does account for why poor performers tend to give an aggrandised estimation of their own achievement, but fails to reveal why those of higher abilities tend to overestimate their accomplishment to a lesser extent, or why the phenomenon is completely missing in the case of top performers.

The level of someone's self-confidence in their judgements also depends on their social circumstances. Social psychologists (e.g., Sherif, Sherif and Nebergall, 1965) have shown that lay people tend to change their judgements about an ambiguous stimulus when paired with someone who is thought to be an expert in the field, or who seems to be very confident in their judgements: lay people's judgements move in the direction of the expert's judgements. Therefore, the expert is negatively influencing their perceptions of the accuracy of their original judgements, and thus their self-confidence in those judgements. Arguably, the judgements entailed in marking a script could involve a lot of ambiguity for a novice marker: such judgements, and a novice marker's self-confidence in those judgements, are therefore vulnerable to the influences of expert markers' comments. Social influences on markers have been investigated in awarding meetings, where candidates' grades are determined by a team of markers using available script evidence (Murphy *et al*., 1995).

Research into metacognition may also explain why poor insight arises. Metacognition has been widely researched since John Flavell first wrote about it in the 1970s (Flavell, 1979). Cognitive skills are seen to be used to solve a problem or task, whereas metacognition is needed to understand *how* a task was solved (Schraw, 1998). A review of the literature reveals that researchers disagree on the nature of the relationship between metacognition and general cognition; some argue that the same cognitive processes are in the background of both problem solving (for example, marking a script) and also of assessing one's own performance in the given task (Davidson and Sternberg, 1998). This would explain why people with lower cognitive abilities tend to overestimate their test performances (Dunning *et al*., 2003). Others (Borkowski, 2000) describe metacognition as a qualitatively distinct executive process which directs other cognitive processes.

Schraw's theory of metacognition (Schraw, 1998) provides a framework which yields alternative explanations for the findings described earlier, and also a background against which markers' experiences, the marking process, providing self-assessment and receiving feedback can all be comfortably placed. Arguably it is the most comprehensive, therefore, our hypotheses and discussion will be based mainly on this theory. According to Schraw (1998), metacognition is said to have two components: *knowledge of cognition* and *regulation of cognition*. Knowledge of cognition includes three different types of metacognitive awareness: declarative awareness, i.e. knowing *about* things; procedural awareness, i.e. knowing *how*; and conditional awareness, i.e. knowing *when*. Regulation of cognition consists of planning, monitoring and evaluation (Schraw, 1998). These are also the features of metacognition that might differentiate between experts and non-experts in any field.

Arguably, experienced (e.g. 'expert') and inexperienced ('graduate') markers are very different in metacognitive terms. Experts should have extensive declarative awareness (subject knowledge) as they have relevant degrees and normally teach the subjects that they mark. Research suggests they use different cognitive marking strategies for different types of candidate responses (Greatorex and Suto, 2005; Suto and Greatorex, *in press*), therefore, expert markers should have procedural knowledge with extensive conditional knowledge as well. Inexperienced graduate markers, by definition, must also have appropriate declarative awareness (subject knowledge). However, they may lack sufficient procedural knowledge (for lack of opportunity to develop and use efficient marking strategies, for example) and therefore are likely to lack conditional metacognitive awareness as well. Apart from their disadvantage in their lack of knowledge of cognition, inexperienced markers may also lack practice in the regulation of cognition, simply because they have never been involved in the planning, monitoring and evaluation features of the marking process. Therefore, inexperienced markers are likely to have considerably weaker metacognitive skills overall, and it could therefore be expected that they will show less insight into their marking.

However, just like any other cognitive skill, metacognition can be enhanced, among other things, by practice, and this in turn can improve performance (in this case, marking accuracy) (Koch, 2001; Dunning *et al*., 2003).

The 'Marking Expertise' research project

The research explained in this article was originally embedded in a major project on marking expertise (Suto and Nádas, 2007a, b, *in press*). The project examined how expertise and various other factors influence the accuracy of marking previous GCSE papers in maths and physics. The main aim was to investigate possible differences in marking accuracy in two types of markers: experts and graduates. For both subjects, the research involved one Principal Examiner, six experienced ('expert') examiners with both teaching and marking experience and six graduates with extensive subject knowledge but lacking marking and teaching experience. All participants were paid to perform question-by-question marking of the same selections of examination questions collated from previous GCSE papers. The experimental maths paper consisted of 20 questions, the physics paper had 13 questions. Stratified sampling methods were used to select candidate responses for each question, which were photocopied and cleaned of 'live' marks. Two response samples were designed for both subjects; a 15-response 'practice' sample and a 50-response 'main' sample for each question. The marking process for each subject was the following: all markers marked the practice sample at home, using mark schemes. They then obtained feedback at a single standardisation meeting led by the appropriate Principal Examiner. The main samples were then distributed and were marked from home, and no feedback was given to markers on the last sample.

The marks of the Principal Examiners were taken as 'correct' or 'true' marks and were the basis for data analysis. Three accuracy measures were used: P_0 (the overall proportion of raw agreement between the Principal Examiner and the marker); Mean Actual Difference (MAcD, indicating whether a marker is on average more lenient or more stringent than his or her Principal Examiner); and Mean Absolute Difference (MAbD, an indication of the average magnitude of mark differences between the marker and the Principal Examiner) (for a discussion of accuracy measures, see Bramley, 2007).

Surprisingly, expert and graduate markers were found to be very similar in their marking accuracy both on the practice sample and on the main sample, according to all three accuracy measures. For maths, out of 20 questions in the practice sample, only three showed significant differences between the two types of markers. On the main sample, a significant difference was found on only one question, where graduates were slightly more lenient than the Principal Examiner and experts. For physics, significant differences arose on three questions (out of 13) on the practice sample and on two questions on the main sample. It is worth noting that despite the significant differences, the graduates also produced high levels of accuracy on all questions. There was some improvement in accuracy from the practice sample to the main sample for both groups. As further data analysis showed, the standardisation meeting and marking practice had a beneficial effect on both groups, benefiting graduates more than experts in both subjects.

Aims and hypotheses of the present study

In a further study within our marking expertise research, which is the focus of the present article, we investigated how markers perceived their own marking performance. Our study of insight and self-confidence entailed administering questionnaires at three points during the marking process, and had multiple aims:

Aim 1: To explore experts' and graduates' self-confidence in their marking accuracy before the standardisation meeting.

According to metacognitive theory, and given that graduates are often assumed to be generally less accurate than experts, two hypotheses are plausible; (1) graduates are aware of their lack of metacognitive skills compared with the experts, and they therefore report a lower level of self-confidence after marking the practice sample; and (2) graduates are not aware of their disadvantage, and all participants' self-confidence levels are very similar after marking the practice sample. The first of these hypotheses would seem most probable, as the graduates were informed at the start of the study that expert markers would also be taking part.

Aim 2: To explore changes in experts' and graduates' self-confidence throughout the marking process.

Metacognitive theory would predict that experts' self-confidence would be high throughout the marking process, and might even show a slight improvement, because more marking practice and feedback on the specific exam questions might develop their metacognitive skills as well. It seems reasonable to hypothesise that graduate markers will report rising levels of self-confidence because they should gain marking experience during the process. Therefore, graduates should report increasing self-confidence on each consecutive questionnaire, even to the extent where their self-confidence level reaches that of the experts.

Alternatively, metacognition theory would suggest that graduates' self-confidence levels will drop on the second questionnaire (after the standardisation meeting), for two reasons; first, graduates' judgements might be influenced by the presence of expert examiners at the standardisation meeting, and although they had known about their involvement in the study, expert examiners might have presented a new frame of reference to which to compare their lack of expertise; secondly, they had just received feedback on the Principal Examiner's 'true' or 'correct' marks, and might have had to reconsider their accuracy on the practice sample regardless of the presence of others. This also predicts that graduates' and experts' self-confidence would be the highest on the main sample, and it will be very similar for the two groups.

Aim 3: To explore the initial pattern of insight of experts and graduates, and see whether there are any significant differences between the groups.

Metacognitive theory would predict that only graduates will show poor insight because they lack procedural and conditional metacognitive awareness, while experts should utilise their previous experience in marking and receiving feedback on their accuracy.

Aim 4: To explore whether participants' insight improves through the marking process.

Metacognitive theory would suggest that all participants, but especially graduates should improve their insight with each consecutive questionnaire, because by that time they will have practised marking as well as received feedback (at the standardisation meeting), and will have practised metacognitive skills by giving account of their insight in our questionnaires.

As mentioned earlier, the literature suggests that some researchers see metacognitive abilities as utilising the very same cognitive processes which are used for the problem-solving task itself; others see it as a superior, organising process of other cognitive processes. Since in the first study in our marking expertise project graduates and expert markers were found to be very similar in their performance of marking accuracy (Suto and Nádas, *in press*), we can assume that it is not their basic cognitive abilities which will discriminate between the metacognitive abilities of the two groups (if we find that these differences indeed exist). If this argument is true, then any difference found in the metacognition of the two types of markers could account for differences in the above-mentioned processes (*procedural awareness*, knowing *how*; and *regulation of cognition*, i.e. planning, monitoring and evaluating), rather than for differences in cognitive skills; this could indicate that metacognition and other cognitive processes are not essentially the same phenomena.

Method

Participants

As mentioned previously, 26 markers were recruited: for each subject, six expert markers (with subject knowledge, experience of marking at least one tier of the selected examination paper, and teaching experience), six graduate markers (with subject knowledge but no marking or teaching experience) and one highly experienced Principal Examiner took part in the study.

Procedure

All markers received a letter at the start of the study, informing them that both expert and graduate markers would be participating in the study, and that all markers would mark the same 'practice' and 'main' samples of candidate responses, on a question-by-question basis. Markers filled in questionnaires on three occasions: (1) at the start of the standardisation meeting, after having marked the practice sample (15 responses) at home; (2) after having attended the standardisation meeting; and finally (3) after marking the main sample (50 responses) at home.

In questionnaires 1 (at the start of the standardisation meeting) and 2 (at the end of the standardisation meeting) each marker was asked:

> *How accurately do you feel you have marked the **first** batch [the practice sample] of candidates' responses?*

In questionnaire 3 (after having marked the main sample), each marker was asked:

> *How accurately do you feel you have marked the **second** batch [the main sample] of candidates' responses?*

To each of these questions, the marker had to circle one of the following answers:

1. Very inaccurately
2. Inaccurately
3. No idea
4. Accurately
5. Very accurately

Results

After checking the distributions of the data, mean self-confidence ratings were calculated and t-tests and Mann-Whitney U-tests were used to analyse possible differences between the two types of markers. Pearson's and Spearman's correlation coefficients were calculated to explore whether there were any relationships between actual marking accuracy and the relevant data on self-confidence.

Analysis of self-confidence of expert and graduate markers

Figure 1 shows the mean self-confidence ratings of expert and graduate maths markers on the three occasions when the questionnaires were administered. According to t-tests, graduates and experts differed significantly in their self-confidence ratings of the practice sample in questionnaires 1 ($t = 4.02$, $p < 0.01$) and 2 ($t = 2.87$, $p < 0.05$), where graduates showed significantly lower confidence in their marking accuracy. This difference disappeared in questionnaire 3 ($t = 1.86$, $p > 0.05$); the two marker groups were similar in their estimations of how accurately they had marked the main sample. Change in self-confidence was only found for the graduates, whose self-confidence improved significantly from the first to the third questionnaire ($t= -3.83$, $p < 0.05$).

Figure 2 shows the mean self-confidence ratings of the physics markers. The ratings of experts and graduates were compared. In contrast with maths, no significant differences were identified between the two marker groups on any of the three questionnaires.

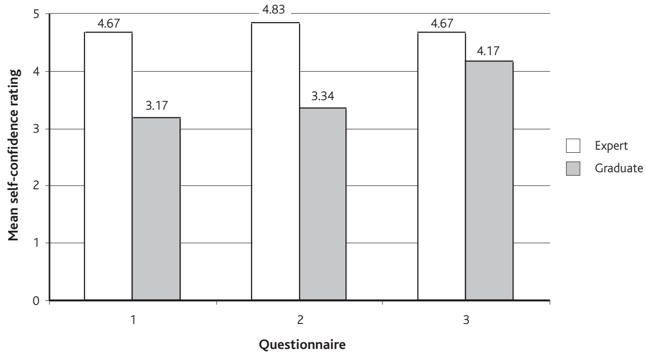

Figure 1 : Graph showing the mean self-confidence ratings of expert and graduate maths markers

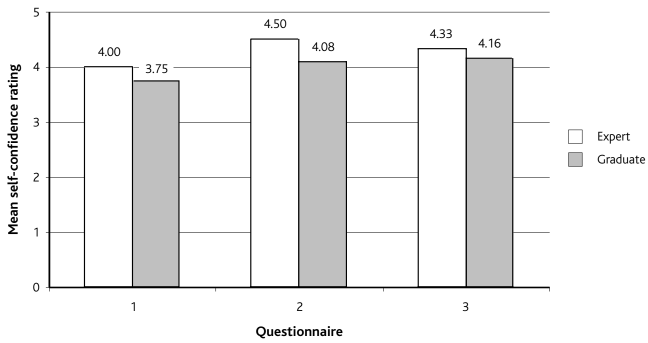

Figure 2 : Graph showing the mean self-confidence ratings of expert and graduate physics markers

Analysis of insight of expert and graduate markers

In order to ascertain whether markers had any insight into their own marking performances, we attempted to correlate the self-confidence data of the two types of markers with their three mean marking accuracy measures (P_0, MAcD, and MAbD) for the practice and main samples.

For maths, Pearson correlation coefficients revealed that neither expert nor graduate markers had real insight into their marking accuracy on either sample; their self-confidence ratings were not significantly related to any of their accuracy measures. The coefficients were the following: for experts: $r = -0.46$, $p = 0.36$ on questionnaire 1; $r = -0.29$, $p = 0.58$ on questionnaire 2; and $r = -0.47$, $p = 0.34$ on questionnaire 3; for graduates: $r = 0.43$, $p = 0.40$ on questionnaire 1; $r = 0.02$, $p = 0.97$ on questionnaire 2; and $r = 0.46$, $p = 0.35$ on questionnaire 3.

For physics, Spearman's and Pearson's correlation coefficients indicated some significant correlations. A significant positive correlation was found for experts' self-confidence after marking the main sample (questionnaire 3) and their mean P_0 values on the main sample ($r = 0.83$, $p < 0.05$) and there was a strong negative correlation with their mean MAbD ($r = -0.86$ $p < 0.05$). Conversely, graduates' self-confidence was significantly negatively correlated to their mean P_0 values ($r = -0.81$, $p < 0.05$) and was positively correlated to mean MAbD values ($r = 0.86$, $p < 0.05$) after the standardisation meeting (on questionnaire 2). Both these correlations indicate that the more accurately the experts marked the main sample, the higher level of self-confidence they reported. Thus, they displayed insight into their own marking accuracies on the main sample. However, the opposite is the case with graduates on the practice sample: the higher self-confidence ratings they gave, the more inaccurate (on two measures) they proved to be. Table 1 summarises the findings.

Table 1: Summary of findings on the correlations between self-confidence levels and marking accuracy

	Does self-confidence on questionnaire 1 correlate significantly with accuracy on the practice sample?	*Does self-confidence on questionnaire 2 correlate significantly with accuracy on the practice sample?*	*Does self-confidence on questionnaire 3 correlate significantly with accuracy on the main sample?*
Maths experts	No	No	No
Maths graduates	No	No	No
Physics experts	No	No	Positive correlation
Physics graduates	No	Negative correlation	No

Discussion

Overall, our results are mixed: our hypotheses were only partially supported by the data, and we found very different patterns of self-confidence and insight for maths and physics markers.

Our first aim was to explore experts' and graduates' self-confidence before the standardisation meeting. All expert markers showed high levels of initial self-confidence; the maths experts' mean level was slightly higher than that of those of both groups of physics markers. It seems that our two hypotheses, namely, that graduates will either report the same level of self-confidence as experts do, or that they will show less self-confidence than that of the experts on the practice sample, applied to one of the graduate groups each: maths graduates showed significantly lower self-confidence than experts, which might reflect expectations of lacking metacognitive and marking skills. Physics graduates, however, showed no difference in their self-confidence from that of experts; in the metacognitive framework this could mean that they did not attempt to account for their lack of experience. However, when these physics graduates' high levels of accuracy are taken into account, their high levels of self-confidence seem only to reflect the expectation of this performance. Finally, it remains a mystery why maths and physics graduates reported different patterns of confidence on the practice sample.

Our second aim was to explore changes in graduates' and experts' self-confidence during the marking process. Metacognitive theory can account for the finding that experts' levels of self-confidence were consistently high; however, no rise was found in their levels of self-confidence over the course of the marking process. Although metacognitive theory would have predicted a small rise, the amount of marking entailed in the study may not have been enough to develop metacognitive skills further. Alternatively, the experts' metacognitive skills may already have been at ceiling level at the start of the research.

As hypothesised, maths graduates were found to report improving levels of self-confidence, up to the point where the significant difference between experts and graduates that had been found previously on the first and second questionnaires disappeared after the main sample had been marked. However, physics graduates were just as confident as experienced examiners were throughout the marking. This is surprising given that graduates, when estimating their own performance, should have taken into consideration their lack of previous marking experience (which they seem to have failed to do on the practice sample already). Nevertheless, they were almost as accurate as experts were, so arguably the equal level of confidence is appropriate but unexpected, as is their high level of marking accuracy.

The data did not support our further hypothesis; the graduates' self-confidence level did not drop after the standardisation meeting in either subject. It seems that the new social reference (expected to be brought about by the presence of experts) or the feedback process did not influence graduates' self-confidence in either subject. However, we did find that all graduates' self-confidence reached the highest level after having marked the main sample, when all previous differences from the experts (if any) diminished.

The third aim was to explore participants' initial insight into their marking accuracy, as indicated by potential correlations between self-confidence and accuracy. Surprisingly, no markers showed any insight on the practice sample before getting feedback at the standardisation meeting. This is especially interesting in the case of expert markers, because metacognitive theory predicts the contrary, counting on their previous experience in evaluating their own marking accuracy. It seems that previous experience in marking different exam questions and in reflecting on one's marking might not generalise to marking new items and to evaluating recent marking accuracy.

Lastly, we explored possible changes in insight in the four marker groups over the course of the marking process. Metacognitive theory would predict that all groups, but especially graduates of both subjects, would improve their insights with each consecutive questionnaire. For maths, surprisingly, neither group showed an improvement in their metacognitive performance with more practice, as neither showed insight on either the practice sample after the meeting, or on the main sample. Data from maths markers, therefore, do not support the metacognitive hypothesis.

For physics, our predictions were, again, only partially supported: experienced markers did show some insight into their marking but only on the main sample. In this case, it seems, the argument that metacognition can be improved by practice was supported by data. Surprisingly, a significant negative correlation was found between physics graduates' estimates and their performance on the practice sample; this, however, seems to support the self-serving bias theory, which predicted this exaggerated optimism. However, the theory predicted the same for all groups, which was not supported by our data.

It has to be noted that because marking was remarkably accurate on the main sample for both experienced and graduate physics markers, we cannot conclude that the difference between their metacognitive abilities is due to different cognitive abilities. Indeed, it may well be that it is the lack of regulation of cognition and procedural knowledge that accounts for different abilities in metacognition. This also sheds light on the nature of the relationship between cognition and metacognition; as graduate physics markers performed similarly to experts on a cognitively demanding task, but they showed a different pattern of metacognition, this suggests that the two processes might not be essentially the very same phenomena. Of course, further empirical research is needed to examine this point in detail.

Limitations

Just as with all research, our study had some limitations. One of the most obvious ones is that the study involved small groups of participants, which did not allow for the detailed analysis of possible age and gender differences in self-confidence and insight. Participants differed from one another on multiple variables; expert markers had both teaching and marking experience, whereas graduate markers were all young professionals. Also, many of the graduates had attended the University of Cambridge, which might have an effect of its own; for example, Cambridge graduates might be more academically focussed; or more or less conscientious or self-assured than graduates from other institutions. A wider variety of expertise and backgrounds of markers is needed for further research.

A further limitation is that the study involved just two examination papers, which were similar in nature. Using other subjects might have produced different outcomes. Another cause for concern is that there is no way of knowing how seriously markers took our questionnaires; whether they took the time and thought about their confidence in their accuracy overall, or whether they just entered a figure without much self-reflection. This uncertainty also stems from the use of an 'experimental' examination process, created for research purposes only, and the marks given had no effect on any candidate's life chances. Had it been 'live' marking, we might have found different levels of self-confidence and insight. And finally, another source of limitation is that marking practice and metacognitive tasks were always performed at the same time, thus the design of the study did not allow for a separate evaluation of effects; a further study would need the separation of these tasks.

Conclusions and further research

Markers of different subjects show very different patterns of self-confidence and insight. Graduate maths markers showed significantly lower self-confidence than maths experts on the practice sample, but not on the main sample. Physics graduates were as confident as expert markers were throughout the marking process. Generally, markers reported constant levels of self-confidence throughout the marking process; only maths graduates improved their self-confidence from the initial marking of the practice sample to the main sample.

Some markers showed some insight into their marking, but this was not consistent, and even experts' insight was not always accurate. Maths markers showed no insight into their accuracies on either the practice or the main sample. Physics experts showed correct insight on the main sample; graduates showed a significant negative correlation between their performance estimates and their actual marking accuracy on the practice sample.

Because of the mixed results, no one theory fully explains all our data; however, it seems that most, but not all of our results can be interpreted in the framework of the theory of metacognition. Thus, this study also serves as an empirical investigation into the nature of the relationship between cognition and metacognition. Differences in insight between experienced and graduate physics markers did not reflect their overall similarity in accuracy; therefore, differences in metacognitive abilities should reflect differences in procedural and conditional awareness, not cognitive abilities. This suggests that cognition and metacognition may entail qualitatively different processes. It is unclear why maths and physics markers showed such different patterns of self-confidence and insight.

As mentioned in the introduction, one practical implication of this study is for standardisation meetings, where the Principal Examiners and their teams discuss questions on which examiners think they were inaccurate. However, the present study has shown that, especially for maths markers, examiners do not have insight into their own accuracy, therefore they cannot tell which questions should be discussed at the meeting. This could be resolved by on-screen marking, where standardisation procedures can entail immediate feedback on marking accuracy, thereby improving markers' insight; or by conducting qualitative studies (using the Kelly's Repertory Grid technique, for example) which invite Principal Examiners as participants to generate further information on what features of a question make it more difficult to mark than others (see Suto and Nádas, 2007b).

Inquiry into markers' metacognition has been extended in an ongoing follow-up study, where several of the limitations of the first study have been eliminated by a more sophisticated research design. In this experimental marking study, we are looking at how over eighty participants with different background experiences mark business studies GCSE and biology International GCSE (IGCSE) examination papers. Markers' metacognition and aspects of their personalities are being investigated using extended questionnaires. The data analysis of this study is currently under way. We are planning to share our results in 2008.

References

Armor, D.A. & Taylor, S.E. (2002). When predictions fail: The dilemma of unrealistic optimism. In: T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgement*. Cambridge: Cambridge University Press, 334–347.

Borkowski, J.G. (2000). *The assessment of executive functioning*. Paper delivered at the annual convention of the American Educational Research Association, New Orleans, April 2000.

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, **4**, 22–28.

Buehler, R., Griffin, D., & Ross, M. (1994). Exploring the 'planning fallacy': Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, **67**, 366–381.

Davidson, J.E. & Sternberg, R. J. (1998). How metacognition helps. In: D.J. Hacker, J. Dunlosky & A.C. Graesser (Eds.), *Metacognition in educational theory and practice*. London: Lawrence Erlbaum Associates, 47–68.

Dunning, D., Meyerowitz, J.A., & Holzberg, A.D. (2002). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments and ability. In: T. Gilovich, D. Griffin & D. Kahneman (Eds.) *Heuristics and biases: The psychology of intuitive judgement*. Cambridge: Cambridge University Press, 324–333.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognise their own incompetence. *Current Directions in Psychological Science*, 83–87.

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 Novakovič 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.