Table A1: Details of the report selection process

Subject area	*Reports placed in each subject area after initial verification of title by Chief Examiner as 'yes/maybe' (N = 118)*	*Reports placed in each subject area after final consideration of titles (N = 94)*	*Reports initially selected for full sample of 20 (N = 23)*	*Reports finally selected for full sample of 20 (N = 20)*	*Reports used in the IRR marking study (N = 20)*	
					Reports selected for the main sub-sample (N = 5)	*Reports selected for the practice sub-sample (N = 15)*
Art & architecture	2	2	1	1	0	1
Biology	11	6	1	1	0	1
Biomedical ethics	11	10	2	2	0	2
Chemistry	2	2	1	1	0	1
Economics	10	8	1	1	0	1
English & applied linguistics	7	5	2	1	0	1
French	4	3	1	1	0	1
Geography	5	5	2	1	0	1
History	6	6	2	1	1	0
Law	8	7	1	1	1	0
Maths & computing	4	4	1	1	1	0
Music, film & drama	7	5	2	2	0	2
Philosophy & religious studies	7	5	1	1	0	1
Physics & astronomy	7	4	1	1	1	0
Politics	9	7	2	2	0	2
Psychology & sociology	18	15	2	2	1	1

IMPACT OF ASSESSMENT

Towards an understanding of the impact of annotations on returned examination scripts

Martin Johnson Research Division **and Stuart Shaw** CIE Research

Introduction

For the past few years awarding bodies in England, Wales and Northern Ireland have been obliged to allow assessment centres and candidates to request to see their examination scripts once they have been marked. Guidelines established by the regulator of qualifications in England, the Office of the Qualifications and Examinations Regulator (Ofqual) in conjunction with the Welsh Assembly Government's Department for Children, Education, Lifelong Learning and Skills (DCELL) and the Northern Ireland Council for Curriculum, Examinations and Assessment (CCEA) outline the steps that qualification awarding bodies need to take to ensure that this accountability function is fulfilled.

According to these documents centres and individual assessment candidates have the right to access marked examination scripts under certain conditions which safeguard issues of candidate data confidentiality. There is little empirical study into practices around scripts returned to centres. It appears intuitive that script requests might be considered as a precursor to a results enquiry but what is less intuitive is whether any other uses are made of these returned scripts.

Returned scripts often include information from examiners about the performance being assessed. As well as the total score given for the performance, additional information is carried in the form of the annotations left on the script by the marking examiner. As far as we know there has been no research into how this information is used by centres or candidates and whether it has any influence on future teaching and learning. Moreover, with current technological developments leading to more scripts being processed in digital formats, it is not clear that this annotation information will continue to be carried on scripts back to centres and candidates in the future. This suggests that research is necessary in order to gather evidence about the potential consequences of such developments and to offer insight into the validity of the inferences that teachers and candidates make about performances based on the annotations that examiners make on scripts.

Literature review

Examiners' annotations have been the subject of a number of recent research studies. Crisp and Johnson (2007) found that examiners' annotations performed two principal functions; communicating the reasons for marking decisions between different members of the assessment hierarchy, and facilitating examiners' thinking processes whilst marking. This second aspect has been pursued further in work by Johnson and Shaw (2008), Johnson and Nádas (2009) and Shaw and Johnson (2009) which consider the role of annotation in assessors' comprehension building practices.

The concept of External Knowledge Representations (EKR) can be employed to describe how annotations work as a tool for both supporting cognition (at an individual level) and distributing cognition (by extending understanding through a linked community). Mislevy *et al*. (2007) conceptualises EKRs as vehicles for discourse, used either by a single individual or among individuals at one point in time or across multiple points in time. They can work by overcoming obstacles to human information processing, for example, through supporting limited working or long-term memory. This conceptualisation also sits comfortably with sociocultural learning theory (e.g. Lave and Wenger, 1991) which considers language to be a central mediating tool for both individual and group understanding. Communities that assemble around shared activity develop particular linguistic forms that have specific characteristics and codes. These linguistic forms are important tools for communication within the community and support coherence. Importantly, these linguistic forms can involve elements (e.g. phrases or words) that are relatively meaningless to those outside of the community.

This sociocultural analysis coheres with an Activity Theoretical perspective (c.f. Engeström, 2001) which seeks to explain the problems that can arise between individuals engaged around a shared activity. Activity Theory suggests that tensions, such as misaligned interpretations, can emerge due to individuals having different roles from each other, each with incumbent purposes, leading them to have different expectations of the tools of the activity. For example, in the case of annotations which are tools for both facilitating and communicating thinking, examiners and teachers might use the same tool but use it differently according to their differing respective purposes. Examiners will tend to work within the rules of the awarding body, which might involve a codified set of annotations that are well understood within a tight community of examiners and which focus on performance summary. Teachers, on the other hand, might prioritise more elaborated annotations which provide a formative function as to what a learner needs to do to improve for a future performance. Whilst both of these perspectives are legitimate and reflect the different purposes that can justifiably be served by annotation tools, they also represent a potential point of conflict.

Since the principal focus of enquiry for earlier annotation studies has been to consider the ways that annotations affect examiners' judgements and communication, it is a natural development to look also at the effect of annotations on non-examiners, for example, candidates and their teachers, who might also have access to the annotations but who were not the intended audience. Although there has been no formal research work, to our knowledge, about how teachers use annotations on returned scripts, such a study would complement earlier research work about annotations in general by considering the wider impact of annotations beyond the immediate annotator and intended recipient, in effect contributing to a 360 degree view of the annotation process.

Research questions

The project had a number of areas of enquiry:

1. How do teachers and centres use annotations?
2. What is the scale of such use?
3. What importance is attached to the annotations?
4. What factors might influence the interpretation of the annotations?

The issue of whether annotations are used validly or invalidly will be explored in the conclusion of the article.

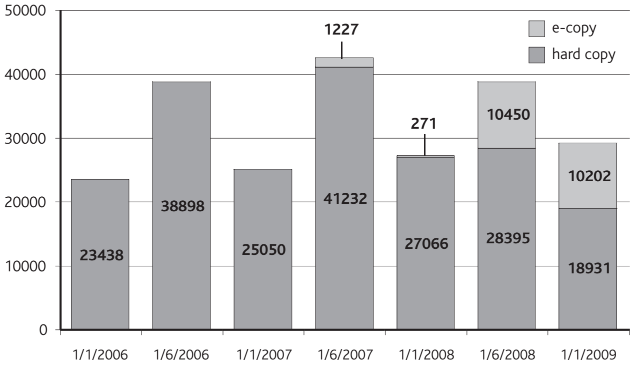
Context for the study

In order to contextualise the findings of the study, data were collected from the OCR script request database to identify any trends in examination script requests from January 2006 to January 2009.

Interrogation of the database suggested that the total number of requests – January and June combined – appeared relatively stable over the 3 years, representing approximately 1% of the scripts processed by OCR each year.

Analysis also shows a growing trend for electronic copies of exam scripts to be returned to centres (Figure 1). This shift reflects the growing numbers of examinations to be digitally scanned for marking purposes and has implications for this particular project since annotations are not typically carried on these scripts.

Close examination of the data from the last full year – 2008 – suggested that the units and centres that accounted for the majority of



*2009 data was collected prior to the closing date for script requests and is therefore partial.

Figure 1: Mode of script request access (2006–2009*)

script requests varied over the two different sessions. For both sessions there is an asymmetrical spread of unit requests, with some units being heavily requested in comparison with mean unit request figures. Similarly, data analysis of those centres requesting scripts showed an asymmetrical balance.

Method

Given the lack of literature related to teachers' interpretation of external examiners' annotations a two-stage research method was adopted. The initial exploratory phase involved semi-structured interviews and focused discussion group sessions with a small group of teachers who shared an in depth understanding of the script request procedure.

Identification of this group of teachers involved an analysis of past script request data. This analysis suggested that English and History teachers might be worthy of inclusion in the study because scripts from these subjects were requested across many different centres and across a variety of units. Psychology was also identified in the analysis because scripts for one of its units were particularly heavily requested by schools.

To identify centres with the greatest use of the script request service a 'measure' was calculated that took into consideration whether a centre had appeared amongst the ten centres that had requested the most scripts following each examination session over a period of three years. Five centres were identified through this analysis, of which four were able to be involved in the initial qualitative interview phase of the project. This involved two English Department heads and two History Department heads from two different schools being interviewed using a semi-structured interview schedule. Furthermore, three Psychology teachers, including two heads of department, from two schools took part in a focus group interview. These meetings took place in January 2009.

During these meetings the teachers were shown a variety of archived scripts from candidates at their own centre. These scripts had originally been awarded marks that fell close to the boundary between two different grades and the teachers were asked about how they might review such performances if requested, and how the annotations on the script might inform these views. The teachers were then asked to assess a script that had been cleaned of all annotations. Following this assessment the examiner annotations were revealed and the teachers were asked to discuss whether their views on the performance were different in light of this additional information.

For the second research stage we reviewed the transcripts and notes taken at the interview sessions and highlighted the main themes that appeared to emerge from the discussions. These themes led to the construction of a survey which aimed to explore the scale of the issues that were identified during the interview and focus group sessions.

These issues included questions about teachers' levels of assessment experience, their script request practices and views on annotations on scripts. 5000 surveys were then distributed to centres who requested script returns between March and June 2009, this number representing roughly one survey for every six script requests in total.

Findings

501 responses (including six empty returns) were returned in the 14 weeks of the script request window, giving a response rate of 10% and a cooperation rate[1] of 99%. Given that the surveys were not posted to any named individual within centres this return/cooperation rate might be considered reasonable, although it is also important to acknowledge the inevitable degree of self selection that relates to remotely administered survey tools.

97% of the teachers responding to the survey (n=448) had requested Level 3 (e.g. A-S/A Level) rather than Level 2 (e.g. GCSE) scripts. Teachers were also most likely to have requested humanities (34%; n=170); science/ electronics/ engineering (27%; n=135); or maths scripts (12%; n=60). These figures are somewhat consistent with those that might have been predicted when considering the returned script profile for the three years prior to the study. 60% of the teachers (n=302) had not examined in the 5 years previous to completing the survey.

Research question 1: How do teachers and centres use annotations?

The interview and focus group data suggested that four uses for requested scripts appeared to be salient for teachers:

- for reviewing exam performances with individual candidates;
- to check that scripts had been marked correctly;
- to use with groups of learners;
- for professional development activities with other teaching staff.

Across the uses there were interesting differences in purpose. The first two elicited uses had an individual focus, with single scripts being used as a tool for review processes and for building an understanding of the characteristics of a particular performance. The second set of uses centred on practices around a range of scripts with a group focus and aimed to support more global understandings about the expected standards of assessment through looking at performances in general.

Research question 2: What is the scale of annotation use?

For reviewing exam performances with individual candidates

The dominant purpose for script requests was to focus on elements of individual student performance. 94% of teachers (n=471) reported using returned scripts to review individual performances, with around 25% of teachers (n=125) systematically using scripts in this way either every session or at least once per year.

Analysis suggested that the primary focus of individual performance review was to inform exam retakes and to maximise candidates' future performance through improving their exam techniques.

To check that scripts had been marked correctly

Requesting scripts to check marking was something that 53% of the teachers (n=263) reported doing, largely on an ad hoc rather than a systematic basis. This practice tended to be instigated by situations where a teacher's expectations about a candidate's performance failed to match the actual exam outcome, leading teachers to request scripts to gain insight into final marking decisions.

Some of this practice appeared to be pragmatic, aimed at using information in returned scripts to question and potentially overturn marks awarded for individual examinees, although it is important not to overstate this view. Whilst some script request practice might be prompted by a teacher's belief that the examination result had under-

[1] This is the proportion of respondents who completed the survey fully. Cooperation rates combine with *response rates* to give a measure of the degree to which a survey is or is not addressing issues that respondents feel to be important.

represented the ability of a particular candidate our data suggest that teachers also tended to use the information from returned scripts for professional development. Rather than taking an initial position of questioning examiner judgement, teachers were likely to be using the scripts to increase their understanding of examiner marking, ultimately in order to align their judgement with that of the examiner through comparing their personal interpretations of the mark scheme with its actual application.

To use with groups of learners

The use of returned scripts with groups of students was reported by 46% of the teachers (n=230), and was considered to be systematic practice for 19% of the teachers (n=95). The primary purpose of this activity was to promote students' understanding of the mark scheme through demonstrating its application and helping to construct a shared understanding of the examiner's view. To do this, teachers tended to use returned scripts to model good performance, often using peer review strategies.

For professional development activities with other teaching staff

Finally, 33% of the teachers reported that they used scripts for professional development purposes (n=165). Comments centred on techniques employed for the purpose of aligning staff perspectives with those expected in examination requirements. This was particularly the case where centres had new department staff. The techniques used with requested scripts tended to involve staff moderation and standardisation sessions which focused on features of good student performance and common errors.

Research question 3: What importance is attached to annotations?

It can be argued that the importance of annotations for those receiving returned scripts relates to the value that they place on those annotations. In turn, we think that the notion of valuing annotations relates to how well the annotations link to the teachers' intended purpose for using those annotations. This is where issues of interpretation and value become intertwined. Different teachers appeared to have different expectations about annotations. The data suggested that these expectations related to whether the teacher had recent examining experience (i.e. within the last 5 years) or not, and that this experience influenced the way that they perceived annotations.

88% of teachers (n=439) agreed that annotations should have a clear link to the mark scheme. When considering perceptions of whether annotations actually did tend to link to the mark scheme only 44% of the teachers (n=222) felt this to be the case, with teachers with current or recent examining experience (teacher-examiners) being significantly more likely than those without examining experience to state that annotations had a clear link to mark schemes (Pearson Chi-square: 8.24769, df=2, p=.016185) (Figure 2).

A key emerging theme throughout the data was the extent to which annotations provided evidence which helped teachers (and candidates) to trust the decisions and judgements of the examiners. 62% of teachers (n=312), regardless of examining experience, agreed that ideally annotations should give information which would help them to trust examiners' judgements.

When looking at reported experience of this phenomenon, examining experience appeared to influence perception levels. Teacher-examiners were significantly more likely than non-examiners to report that annotations did actually reinforce their trust in other examiners' decisions (Pearson Chi-square: 9.40594, df=2, p=.009070) (Figure 3).

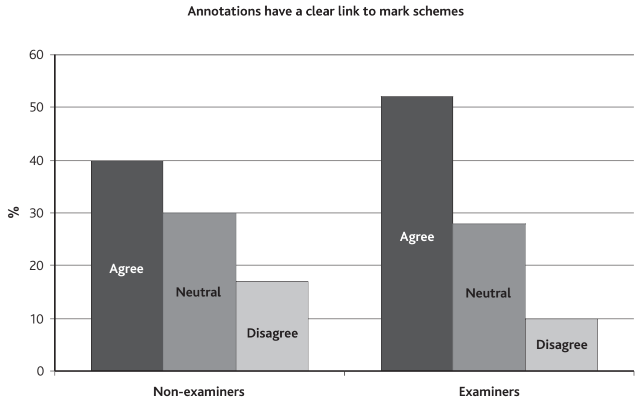


Figure 2: Perceived links between annotations and mark schemes

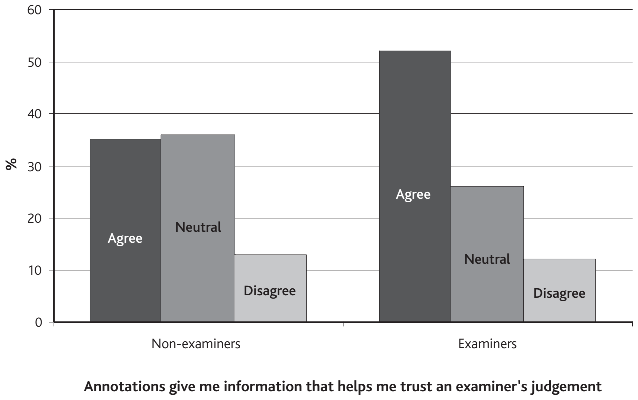


Figure 3: Annotations aid trust

Research question 4: What factors might influence the interpretation of the annotations?

A key theme emerging from the quantitative data was that examining experience appeared to influence the way that teachers were perceiving annotations. Teacher-examiners were more likely to perceive that annotations tended to reflect mark schemes and at the same time give them information which helped them to trust the judgements of other examiners. Further analysis of the qualitative data suggested at least four ways that experience might influence perception of annotations.

Abbreviations:

Teachers suggested that examining experience gave them a greater awareness of the annotation abbreviations that they encountered on returned scripts, for example, '*You know what the abbreviations mean and where you would expect to find them*'. Significantly, this knowledge of abbreviated terms was not in itself of central importance to the teacher-examiners.

Understanding mark schemes:

The most frequently expressed comment related to how examining experience gave teacher-examiners a good understanding of the mark scheme, helping to support their interpretation of other examiners' marking. Importantly, this interpretation relied on them attending to examiners' annotations, for example, '*I have an experienced understanding of mark schemes and how they are applied en masse to students' exam scripts. I understand the shorthand used*'. There was an important difference in the perception as to whether annotations might be seen to illuminate the mark scheme or vice versa. This issue related to teachers' existing levels of mark scheme knowledge, with teachers sometimes making it clear that they had gaps in their mark scheme understanding which they used exam annotations to help overcome. This is an important distinction; whereas examiners tended to describe how they could make sense of annotations in light of their good mark scheme understanding, non-examiners tended to look to the annotations to help them construct their sense of the mark scheme.

Privileged knowledge about assessment:

Communities of practice perspectives (c.f. Lave and Wenger, 1991) suggest that aspects of mark schemes will remain opaque and involvement with a community of assessment practice allows its members to build understandings that are coupled to their experience levels.

Sociocultural perspectives suggest that community members have access to privileged information or 'insider knowledge' through a shared language which links to their involvement in a community of assessment practitioners. This 'insider knowledge' of assessment, through examiners' engagement with other examiners in formal assessment activity (e.g. participation in training and standardisation sessions) not only helps examiners to understand how potentially opaque criteria might be applied in context, but it also allows them insight into the limits to which annotations as tools can fully illuminate the meanings of examiners' judgements in relation to mark schemes. This aspect of comprehension is most clearly expressed by teacher-examiners who highlight some of the nuances of interpreting annotations. Their comments suggest that examining experience helped them to consider meanings that were merely implied by annotations, for example, *'[Examining experience] helps in understanding the relationship between informal marks on the page and the actual mark or part mark awarded for a question',* and, *'I understand what [annotations] imply as well as mean'*.

Recognising the main purpose of annotations is to support the process of the annotator making good judgements:

Examining experience also influenced teachers expectations about the scope of the functions that annotations could be expected to support. There was a significant difference between teacher-examiners' and non-examiners' aspirations that annotations should have a formative function (Pearson Chi-square: 12.0894, df=2, p=.002371). Most non-examiners felt that annotations should highlight where and perhaps how performances might be improved, whilst this sentiment was held by only a minority of teacher-examiners (Figure 4).

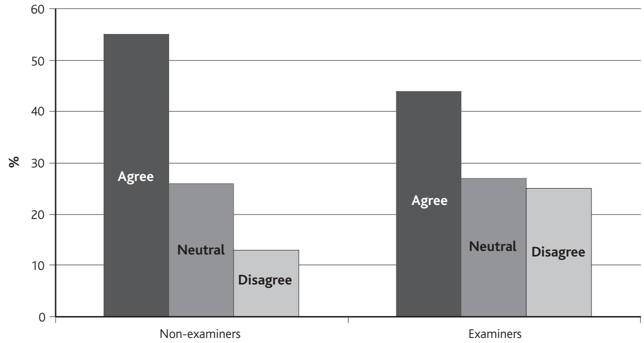


Figure 4: Annotations and formative purpose

This difference in expectation appears to be underpinned by a difference in understanding about the primary purpose of annotating when marking. Whilst annotations are a tool that can help to satisfy the function of providing formative feedback on performances, examiners appeared to be more aware that the primary foci of annotating whilst marking were (a) supporting the examiner's own thinking, and (b) accounting for that thinking to others who have an interest. It is a real concern that if the demands placed on annotating practice stretch beyond these primary functions, for example, to satisfy formative functions, it is possible that the tool itself might fail to support the primary purpose.

Conclusions

One aspect of validity that we have chosen to focus on in this study is 'the extent to which the inferences which are made on the basis of the outcomes are meaningful, useful and appropriate' (Cambridge Assessment, 2009, p.8). This resonates with the view of validity outlined in the Standards for Educational and Psychological Testing (1999):

> *Validity logically begins with an explicit statement of the proposed interpretation of test scores along with a rationale for the relevance of the interpretation to the proposed use. (1999, p.9)*

In our view, annotations have a direct link with validity through the way that they can connect a score, the interpretation of the score, and any ensuing actions based on such an interpretation. The data from this study suggest that important aspects of interpretation are linked to experience within an assessment community of practice. Crisp and Johnson (2007) note that:

> *Despite room for marker idiosyncrasy the key underpinning feature of annotation use appeared to be that it needed to be commonly understood by other members of the community… Situated Learning Theory suggests that effective working communities are based around sets of common norms and practices. Effective communication between community members is essential to the efficient working of the group, and part of this communication might involve the evolution and use of specialised tools which facilitate the transmission of knowledge between community members. To some extent it appears that marker annotation practices conform to this model, behaving as communicative tools and carrying a great deal of meaning to those within the community. (2007, p.960).*

It appears from the present study data that this particular community of practice comprises other examiners and teachers with recent examining experience, and that this involvement through standardisation and training sessions allows a special insight into the interpretation of annotations.

Teachers were more likely than examiners to use annotations to help them to increase their understanding of the mark scheme through looking at how annotations implied the application of marking criteria. This inductive reasoning (inducing the universal from the particular) contrasts with teacher-examiner processes that tended to use generalised mark scheme understanding to interpret the potential meanings of particular annotations. The potential problem with the inductive approach to annotation use is that there is an assumption that the annotations give a 'true' reflection of mark scheme application.

Annotations should not always be expected to carry a clear communicative function due to the fact that they might represent the fluid thoughts of an examiner at a point in time during decision making, containing tacit features that support examiner thinking, and leading to them being difficult to infer meaning from. It is clear that these characteristics could limit the ability of someone to use the annotations at face value to make valid inferences about an assessed performance.

Teachers were more likely than teacher-examiners to expect annotations to provide information that could be used for formative purposes (e.g. showing explicitly where a performance could be improved). This difference in perspective is potentially important since it affects the degree to which annotations should be expected to function as tools to support transparent communication. Since examiner annotations are primarily concerned with the functions of supporting examiner thinking and communicating the reasoning behind a judgement, formative annotating is an extraneous purpose which would possibly confound the primary function of the activity and would therefore be inadvisable. In order to mitigate potentially invalid actions based on script annotations, it is advisable that teachers and candidates are informed about why it would be inappropriate for examiners to make formative annotations on scripts.

Despite the inevitably individualised characteristics of examiner annotations there is still scope for the meanings of annotations to be made more explicit to those who have access to them. This is as true for examiners who are engaged in marking a particular examination paper as it is for the teachers who can read the annotations when they access requested scripts. The inclusion of abbreviated annotation terms and shared meanings might be a useful addition to mark schemes but it is very important to recognise that this is only of superficial importance compared with the insights gained from annotations when teachers have a deep understanding of the mark scheme.

This project contributes to a growing understanding of how annotations function and suggests that the primary concern should be that annotation use be fit for purpose. Whilst validity requires that information relating to an assessment is as transparent as possible, and annotations can assist in this process, it is also important to make the limits of annotations explicit to those who receive them on returned scripts.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC.: American Educational Research Association.

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

Crisp, V. & Johnson, M. (2007). The use of annotations in examination marking: opening a window into markers' minds. *British Educational Research Journal*, **33**, 6, 943–961.

Engeström, Y. (2001). Expansive Learning at Work: toward an activity theoretical reconceptualization. *Journal of Education and Work*, **14**, 1, 133–156.

Johnson, M. & Nádas, R. (2009). Marginalised behaviour: digital annotations, spatial encoding and the implications for reading comprehension. *Learning, Media and Technology*, **34**, 4, 323–336.

Johnson, M. & Shaw, S. (2008). Annotating to comprehend: a marginalised activity? *Research Matters: A Cambridge Assessment Publication*, **6**, 19–24.

Lave, J. & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.

Mislevy, R.J., Behrens, J.T., Bennett, R.E., Demark, S.F., Frezzo, D.C., Levy, R., Robinson, D.H., Rutstein, D.W., Shute, V.J., Stanley, K. & Winters, F.I. (2007). *On the roles of external knowledge representations in assessment design*. University of Maryland: National Center for Research on Evaluation, Standards, and Student Testing.

Shaw, S. & Johnson, M. (2009). *Annotating on-screen: the influence of reading environment on annotative practice and assessor comprehension building*. A paper presented at the International Association for Educational Assessment Annual Conference, Brisbane, September.

Must examiners meet in order to standardise their marking? An experiment with new and experienced examiners of GCE AS Psychology

Nicholas Raikes, Jane Fidler and Tim Gill Research Division

This article is based on a paper presented to the annual conference of the British Educational Research Association held in Manchester, UK, in September 2009.

Summary

When high stakes examinations are marked by a panel of examiners, the examiners must be standardised so that candidates are not advantaged or disadvantaged according to which examiner marks their work.

It is common practice for awarding bodies' standardisation processes to include a 'standardisation' or 'co-ordination' meeting, where all examiners meet to be briefed by the Principal Examiner and to discuss the application of the mark scheme in relation to specific examples of candidates' work. Research into the effectiveness of standardisation meetings has cast doubt on their usefulness, however, at least for experienced examiners.

In the present study we addressed the following research questions:

1. What is the effect on marking accuracy of including a face-to-face meeting as part of an examiner standardisation process?