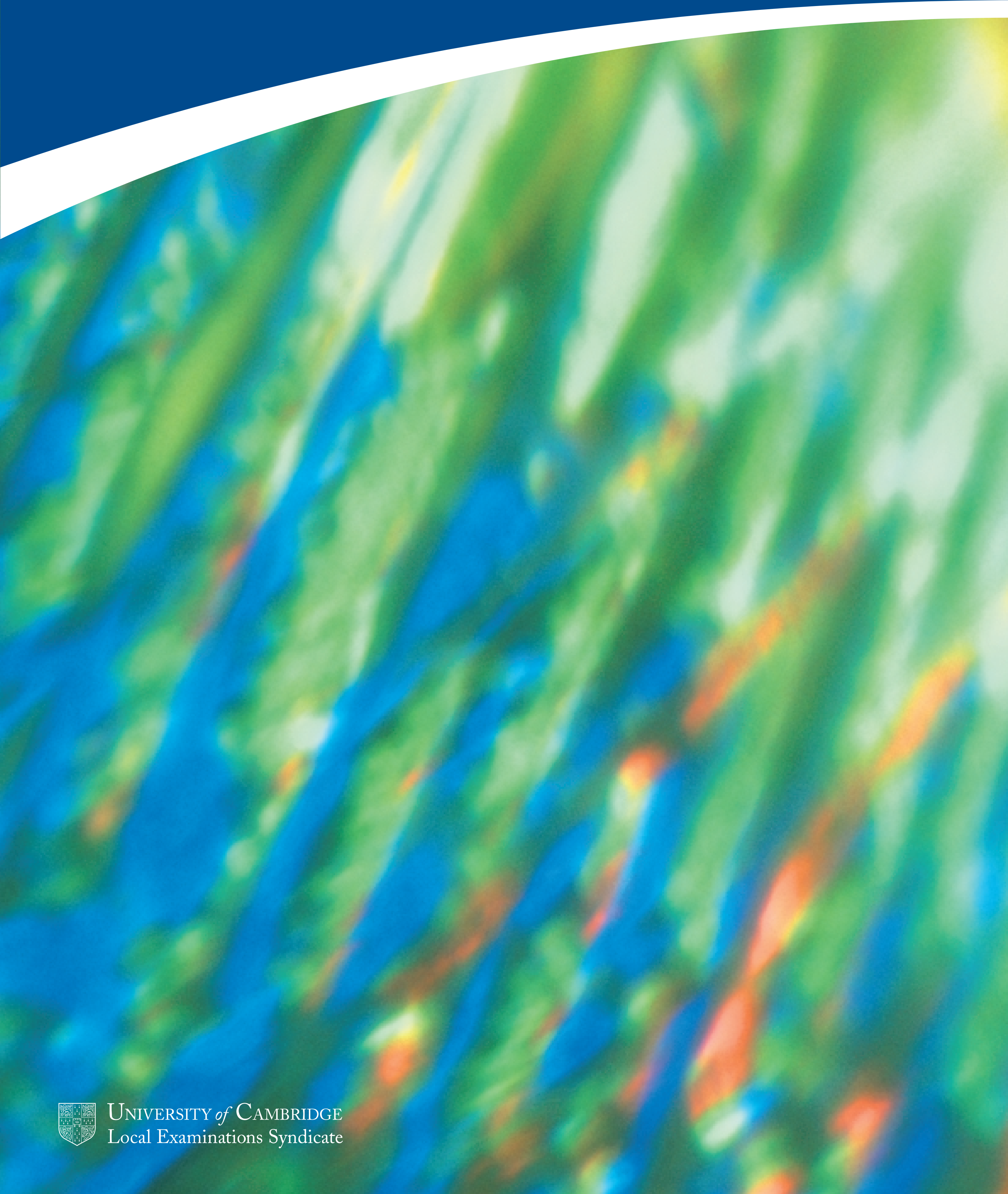
Issue 10 June 2010

Research Matters

CAMBRIDGE ASSESSMENT



UNIVERSITY *of* CAMBRIDGE
Local Examinations Syndicate

Research Matters : 10

A CAMBRIDGE ASSESSMENT PUBLICATION

If you would like to comment on any of the articles in this issue, please contact Sylvia Green.
Email:
researchprogrammes@cambridgeassessment.org.uk

The full issue and previous issues are available on our website:
www.cambridgeassessment.org.uk/ca/Our_Services/Research

Foreword

While the articles in this edition of *Research Matters* again engage with highly analytic approaches to the understanding of the behaviour of specific assessments, a key issue shines from two pieces (see Beth Black's article and the one by Irenka Suto and Stuart Shaw). Much is made in national comment here – and international comparative work across the globe – about the tendency to narrow teaching and learning by focussing on 'that which is easy to assess'. I think that there are healthy contra-indications. In the United States, there continues to be considerable growth and interest in Advanced Placement qualifications – which are highly curriculum-based examinations resembling A levels – punching a hole in the common myth that the multiple-choice SAT reigns supreme in US HE entry. The extraordinary interest in England in Critical Thinking (see Black) and globally in CIE's 'Independent Research Report' (see Suto and Shaw) suggests that educationally-valuable assessments which are nonetheless highly demanding in terms of assessment administration and operation are not universally in decline – and are in fact alive and well. From this we can take a measure of comfort that – in some domains at least – educational value continues to be placed ahead of administrative convenience and drift towards more conservative, 'safe' assessment. Long may this continue.

Tim Oates *Group Director, Assessment Research and Development*

Editorial

This issue covers a wide range of themes including e-assessment, Critical Thinking, quality assurance and methods for studying comparability in vocational contexts. The variety illustrates the depth and breadth of research interests currently under investigation in relation to processes, technological developments and the assessment of new qualifications.

The first two articles concentrate on Critical Thinking. Beth Black reports on the introduction of AS level Critical Thinking for which the candidature has risen dramatically while grades have remained relatively low. This is followed by an article by Joe Chislett, a senior examiner in Critical Thinking. He provides an interesting account of a seminar organised by Cambridge Assessment on the role and value of Critical Thinking.

Suto and Shaw then consider the challenges of marking research reports written by students preparing for university. There are a number of challenges which arise when assessment schemes are designed to reward generic research skills rather than particular subject knowledge. Johnson and Shaw investigate the impact of annotations on teachers and candidates. Their research considers the effects of comments that examiners make on scripts, given that for the past few years centres and candidates have been able to request to see their examination scripts once they have been marked.

Three articles focus on quality assurance in assessment. Raikes, Fidler and Gill report on an experimental standardisation study and ask whether face to face standardisation affects marking accuracy; whether effects vary according to question type and/or the experience of the examiners; and to what extent examiners carry forward standardisation on one set of questions to a similar set of questions. Their work poses some interesting questions for awarding bodies about how they organise their procedures. The second article is a literature review on item level marker agreement. Curcin concentrates mainly on the inter-marker agreement aspect of marking reliability in the context of on-screen marking. She discusses the implications for marking monitoring research and practice in this very topical and challenging area. In the second literature review Matt Haigh examines the evidence around the claims made for the shift towards computer-based assessment (CBA) in educational settings. He highlights some important considerations for researchers undertaking empirical work on CBA in the future.

The final two articles outline the development of new research methods. Greatorex and Rushton investigate the use of a scale of cognitive demands, known as CRAS, which was developed using academic qualifications and ask whether it is suitable for use in comparability studies involving vocational qualifications. In their work on validity Shaw and Crisp address the difficulties of providing validity evidence to support the claims made about assessments. Their research aims to design a set of methods for validation that can be used routinely as part of an ongoing validation programme.

Sylvia Green *Director of Research*

"It's not like teaching other subjects" – the challenges of introducing Critical Thinking AS level in England

Beth Black Research Division

Introduction

This article focuses on the introduction of Critical Thinking AS level into schools in England. As an AS level, it is a qualification, with a specification (syllabus) which prescribes the content that will be examined. As such, it does not itself provide a 'programme of instruction'. Even so, it has probably been the catalyst for the largest scale introduction of Critical Thinking into schools in England. In 2001, 130 schools entered in total just over 2,000 candidates for the whole AS level. By 2009, this had increased to over 1000 schools[1] entering over 22,000[2] candidates.

However, candidate 'success' at Critical Thinking (in terms of proportion of grade As and passes) has remained relatively low, as shown in Figure 1.

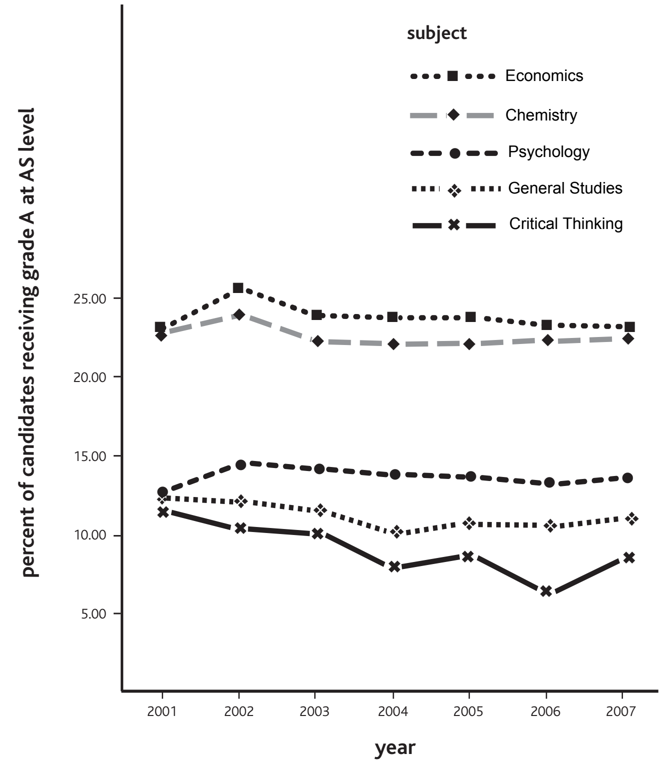


Figure 1: proportion of candidates who have received a grade A in various AS subjects between 2001 and 2007

Hypotheses or potential explanations for why this may be the case fall broadly into three types (though these are not mutually exclusive or independent of one another):

1. Performance standards exhibited by candidates reflect low level of teaching provision.
2. Performance standards exhibited by candidates reflect low level of candidate motivation to achieve in this discipline.
3. The high demand of the discipline.

This article explores the first two types of explanation in detail, but first we briefly consider the third one. The perception that Critical Thinking (CT) is difficult is partly because of the relatively low proportion of candidates, in comparison with other subjects, who receive a grade A as in Figure 1 (this relates to the notion of 'grading difficulty'). While qualification outcomes in terms of % of candidates at or above particular grades may affect the *perception* of difficulty, they do not necessarily mean that a subject is 'easy' or 'difficult'. There is a distinction to be made between *grading difficulty* and *intrinsic difficulty*. But is there a case to be made that, intrinsically, Critical Thinking is difficult? From a cognitive point of view, it might be considered more difficult than some subjects since it necessarily requires abstract, rather than concrete, thinking. For instance, conceptualising the subject in terms of Bloom's taxonomy (Bloom, 1956), it could be said Critical Thinking is characterised by higher-order processes such as evaluation, analysis, synthesis and application.

In terms of hypothesis 2, there is some evidence to support this in Vidal Rodeiro's (2007) survey of students' A level choices. Over 6000 students were surveyed about factors affecting their AS and A level choices. Of the students surveyed, 5.1% were taking Critical Thinking AS level, and nearly half of the centres in the study offered Critical Thinking AS level. While Chemistry, Mathematics, and English were seen as some of the most important subjects at AS level, Critical Thinking was seen as one of the least important (above only Citizenship and General Studies). Of the students taking Critical Thinking, nearly 58.9% rated it as their least important AS subject, while only 8.9% rated it as their most important subject. While this research does not allow us to understand the finer details of how students may have construed the term 'important' (e.g. 'important for me to get a good grade'/'important in terms of the significance of its subject matter' etc.), it does indicate that there is likely to be lower motivation in this subject than in others. This current research project also aimed to find out about student motivation – though only indirectly – through the reports of their teachers.

In any subject, there are undoubtedly difficulties in delivering something new. As a school subject, Critical Thinking has grown so rapidly that few current teachers are likely to be in a position to have the benefit of years of teaching experience in the subject. In the US, where Critical Thinking programmes of instruction have been around for much longer, Sternberg (1987) identified 'eight easy ways to fail before you begin' – eight 'fallacies', peculiar to Critical Thinking, which 'obstruct the teaching of critical thinking … and make it easy to fail'. These include the following (sometimes interrelated) 'fallacies':

- teachers who assume they have nothing to learn from students, whereas teachers need themselves to be receptive to new ideas;

[1] About a third of centres offering AS/A levels in England enter candidates for Critical Thinking.

[2] For context, some other AS subject entries for 2009: Chemistry 58,473, Economics 27,714, French 19,122, English 107,124.

- that what really counts is the right answer – in Critical Thinking it is the thinking behind the answer which is important;
- that the job of a course in Critical Thinking is to teach Critical Thinking – Sternberg's point here is that students and teachers both have to think for themselves and thus the role of the teacher in the classroom is more of a facilitator than a didact.

Sternberg concludes that teaching Critical Thinking, though possible and desirable, is not simple. Some common ideas that teachers hold about teaching and learning, while they may be applicable in the normal course of classroom events, do not apply in the Critical Thinking classroom. This suggests that the struggle of introducing Critical Thinking is not just that of introducing a new subject – it requires a fundamental re-orientation prior to teaching.

This point resonates with the findings of a case study of the implementation of three different Thinking Skills programmes in a UK context (Baumfield and Oberski, 1998). Although they found that the instigation of the programmes was a response to dissatisfaction with the prevailing mode of teaching and learning, it was difficult for teachers to entirely shake off that prevailing mode. The programmes afforded greater opportunities for group work and discussion (and this was seen as important by both students and teachers), but some teachers found this sometimes difficult to manage. In particular, not having a "solid body of content" and trying out new ideas for the first time created some insecurity. There was a tendency at times, therefore, to resort to more familiar modes of teaching in terms of what a productive and meaningful lesson should be. Baumfield and Oberski conclude:

> *… if conventional modes of planning using aims/objectives/outcomes are based, albeit loosely, on behaviourist models of learning, we would anticipate some incompatibility with the constructivist orientation of thinking skills programmes*.

Both Sternberg's and Baumfield and Oberski's papers suggest that the introduction of Thinking Skills/Critical Thinking as a new subject may be more problematic than other new subjects.

Richardson (2008) describes some of the difficulties of introducing another new subject into schools – Citizenship. This possibly has some shared issues with that of Critical Thinking: for example, difficulties in defining the construct (Black, 2008; Kerr and Cleaver, 2004); the teaching and assessment was adversely affected by lack of time, resources and training (House of Commons Select Committee, 2007, cited in Richardson, 2008). This too is potentially an issue for Critical Thinking – certainly, no teacher has a degree in Critical Thinking, no teacher training qualification in the UK includes Critical Thinking. In the US, there is a growing bank of evidence that teacher training in advance of teaching Critical Thinking skills has a significant impact upon the success of the programmes in terms of student gains in Critical Thinking skills (Abrami *et al*., 2008).

The present research considered the difficulties of introducing a new subject through an exploration of the results of a survey of 236 teachers of Critical Thinking and reports on the ways in which centres have implemented the provision of this new school subject.

Method

Sample

As the main medium for collecting data was an electronic questionnaire, we attempted to contact all centres (n=1096) with entries for OCR AS level Critical Thinking units in the June 2007 session by email. There were 236 responses from teachers, representing just over 20% of all OCR Critical Thinking centres and 34.3% of AS candidate entries for 2007. In general terms, this represents a good response rate.

The questionnaire

Prior to the main data collection stage, the questionnaire had been piloted in two stages (involving 5 centres) in order to minimise ambiguities and maximise information capture.

The questionnaire was available for online completion. For those that requested, a paper version of the questionnaire was made available. The majority (n=226) opted for completing the electronic version.

The questionnaire consisted of 50 questions, a mixture of closed and open format questions divided into subsections, as used in the reporting of the results below.

Results

Section A: Background information of respondents

Respondents were overall very experienced teachers (mean teaching years = 18), though, given the newness of Critical Thinking, it was unsurprising they had not been teaching CT for long (mean years teaching Critical Thinking = 2.95).

Figure 2 shows the respondents' main (first) subjects (i.e. the greatest amount of teaching/contact time), second and third subjects. The respondents came from a variety of subject backgrounds. Teachers of Religious Studies/Philosophy, English, History and Science accounted together for more than 50% of the respondents.

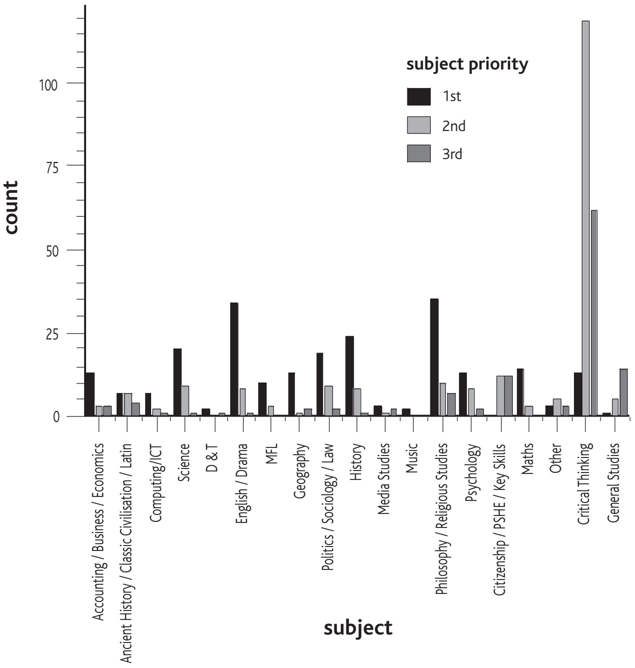


Figure 2: Frequency bar chart showing respondents' first, second and 3rd subjects

We can also see that Critical Thinking was rarely cited as their main subject, but much more frequently as a second or third subject. An analysis of contact teaching hours for Critical Thinking as a proportion of overall hours revealed that more than two thirds of respondents reported that Critical Thinking constituted less than 20% of their teaching timetable. It seems likely that teachers may find it difficult to prioritise Critical Thinking when it forms such a small part of their timetable.

Section B: Timetabling and delivery time of Critical Thinking

The number of hours timetabled in order to deliver the AS programme to students gives an indication of the centres' commitment to Critical Thinking.

Figure 3 shows a wide range of length of programme in terms of contact teaching hours (guided learning hours or glh) for the AS course, ranging from 16 hours to 165 hours. The mean CT AS programme was delivered in 57.12 hours, equivalent to an average of 1.5 hours per week. To provide some context, AS specifications in general suggest that they require approx 140–160 guided learning hours and Critical Thinking is no exception. Thus, it seems that in the majority of centres, Critical Thinking provision was very much attenuated. Teachers were also asked to rate the adequacy of this time. None indicated that they had 'far too much time'; the modal response was 'about the right amount of time' (60%), while 40% of respondents thought that there was not enough time ('too little' or 'far too little time') per class for the delivery of Critical Thinking.

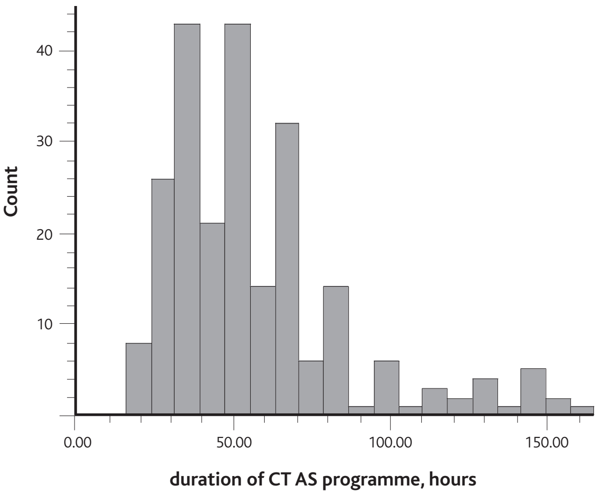


Figure 3: Histogram – programme length of Critical Thinking AS courses in hours

Section C: Student motivation

In brief, teacher reports of student engagement and interest, and attendance were mainly positive (see Figure 4). Attendance can be viewed as a behavioural measure of motivation.

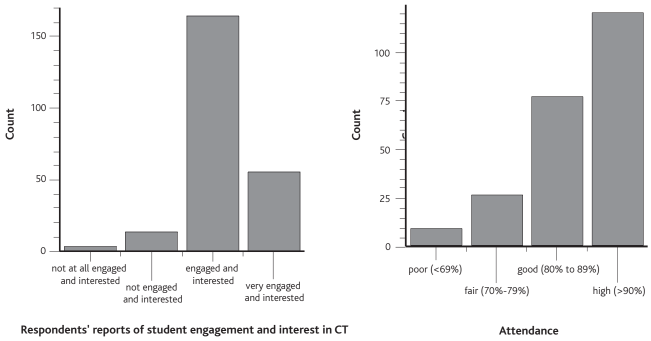


Figure 4: respondents' reports of student motivation

Respondents had the opportunity to write further comments about students' attitudes and motivation towards Critical Thinking. Over half of the respondents added a comment, and these were categorised in the first place as either 'wholly positive' (27.5%), 'wholly negative' (13.3%) or 'positive and negative' (51.7%) in terms of student motivation. The comments provide insight into the delivery of Critical Thinking and student motivation and so it is worth considering them in some depth.

Many of the positive comments made a point about general enjoyment of the lessons/subject (n=31).

> *Students enjoy the fact that this is a different approach to learning from the majority of their subjects.*

> *Students seem to value the subject as being different; they enjoy the immediacy of its challenges; they enjoy the way it enhances their ability to win arguments … they think it's 'cool'.*

Some teachers (n=9) reported that students could see the benefit (transferability) to their other A levels/GCSEs:

> *Students enjoyed the course. While many found it intellectually challenging and may come out with low grades there was a real sense of achievement for completing the year. All students felt it helped them in other subject areas.*

However, a number of positive responses about student motivation were tempered with other issues. Some of these themes are reported below. One prevalent theme was the priority students gave to Critical Thinking when demands of other A levels (or GCSEs) increase (n=33).

> *When taken as an extra there are always problems around pressure times such as coursework deadlines.*

A number of responses (n=13) noted the impact of timetabling, for example, lessons timetabled outside of normal teaching hours or simply not having enough time. Responses seem to indicate that such practices can effectively sabotage the course.

> *Attendance was an issue for a significant minority in that it [clashed with] other activities students committed to, e.g. rehearsals for stage productions or rugby trials etc.*

For some teachers, the mandatory nature of the course and removing student's ability to opt for the subject, had a negative effect on motivation (n=9).

> *Whilst the majority of the students see it as beneficial, there are some that resent having to do a compulsory subject once they are in the Sixth Form …*

More rarely reported were problems with the motivation of the teachers (n=4) who themselves may have had Critical Thinking imposed upon them like their students.

> *As they [students] don't choose this subject as a main AS they give the subject very low priority. Interest in the subject depends very much on who is teaching it. Often staff who are uninterested in the subject are asked to teach CT to fill their timetable.*

Perceived subject difficulty was mentioned in a number of responses (n=23), sometimes along with its impact upon motivation and/or attendance or retention.

> *Attendance is much better than for other enrichment options. Students recognise its value but worry about the effect it has on their grade profile as they are used to getting grade As.*

> *Our students are mainly motivated by the possibility of top grades. In a high achieving school like ours they may be discouraged from doing CT in case they get a B or lower; this would be a 'stain' on their record. Only the very top students welcome the challenge (sad but true).*

The latter comment is quite interesting in motivational terms. For some students, acquiring Critical Thinking skills has no intrinsic value, is not an end in itself, but is only worth persevering with if it were a means to some other end. In this case, getting a top grade as an outcome is the (only) incentive.

Following on from this, a number of respondents commented upon the perceived value of Critical Thinking, frequently identified in terms of UCAS points, though not always.

> *They ... don't see the value as it does not count towards many university applications.*

There seems to be strong picture emerging that Critical Thinking, though with much potential to be rewarding and engaging, is often faced with difficulties that affect motivation.

Section D: resources and training for teachers of Critical Thinking

The majority of teachers (92%) had attended at least some specific CT training events and many could identify an aspect of their formal academic education that had helped them teach Critical Thinking (52%).

In terms of the latter, respondents often reported that a degree (or part of a degree) in logic and/or philosophy had been useful (see Figure 2 earlier – many Critical Thinking teachers are primarily Philosophy or Religious Studies teachers). However, there was a wide range of responses across the range of arts, humanities and science domains. A few respondents elaborated on how their degree had helped, showing that some teachers have been able to see *how* Critical Thinking skills are embedded within their own education and the structure of a particular discipline. For instance:

> *Psychology involves a critical approach to both data and written argument.*
>
> *Theology degree – many units considered the nature of arguments, concept of proofs etc.*
>
> *Economics – the analytical requirement of the subject.*
>
> *Mathematician – naturally logical!!*

The most common form of training attended (71.2% of respondents) was awarding body INSET[3] and many had attended INSET from other providers. Many reported partaking in other (less formal) types of training such as discussions and ideas sharing with other teachers, either within their centre or in a local network. The overall picture, though, was that most teachers had only experienced a handful of relatively brief training experiences. They were largely self-taught and had had to be largely self-reliant. This was evidenced in many of the teachers' comments:

> *... I do have the opportunity to attend inset but my other subjects and classes take priority and I have not yet felt I can fit in a training session for myself.*
>
> *Perhaps the most useful training has been in my role as an Assistant Examiner. This has enabled me to develop an excellent understanding of what is required by students in order for them to achieve top grades.*
>
> *None. I've done it all by myself!*
>
> *I had to pay for course myself.*
>
> *I feel there is a desperate need for far more training for people who like me are 'flung in at the deep end' and have little clue of what they are expected to deliver! I have had one useful day of training with a trainer brought in by our new Head Teacher who recognised the lunacy of what was happening (i.e. go teach this with no training) and one day which was really too advanced for where I was at the time.*

Section E: Other questions

This final section of the questionnaire asked questions aimed to investigate wider attitudes towards Critical Thinking, both their own attitudes and their students'.

Teachers were asked whether they encouraged students to apply Critical Thinking techniques or think more critically in the other subjects that they taught. 92.3% responded that they did so at least sometimes, with many providing interesting additional comments which typically referred to multiple Critical Thinking skills (argument, analysis, evaluation, consideration of bias/credibility etc.). Some of these comments are included in order to illustrate how teachers have found that Critical Thinking skills can have a useful application in different subject domains.

> *In my science lessons when considering social impact/consequences of things – e.g. genetic medication or choosing the location of chemical plants.*
>
> *When listening/reading a text in French AS/A2 we approach it from critical thinking perspective of key purpose, reasons used, assumptions made, inferences drawn etc.*
>
> *[CT] models the type of reasoning they need to use in their own essay writing.*
>
> *In English: to think about their arguments in essays and the ways in which they present their views trying to provide strong evidence/reasons to back up their conclusions.*
>
> *In sociology I highlight types of flaws in arguments; I always encourage [students] to structure arguments carefully.*

Respondents were asked about whether they and their students value Critical Thinking. Results are presented in Figure 5.

Both graphs show overall that both teachers and students (according to teachers' reports) tend to positively value CT. And although the teachers' graph is *more* positive, this would probably be true of any subject. On the whole, teachers said that they highly value Critical Thinking, frequently backed up by additional comments showing great enthusiasm for the subject:

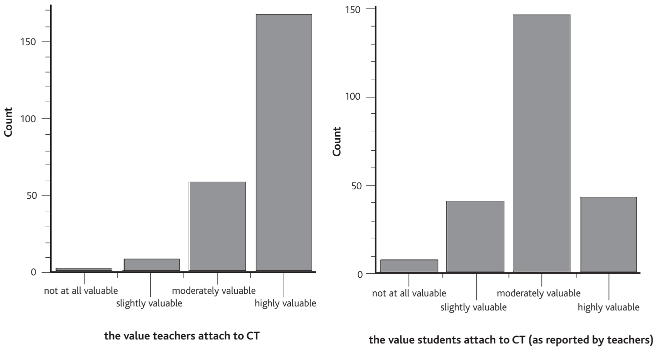


Figure 5: value attached to CT by both teachers and students (as reported by teachers)

[3] INSET means 'In-service training' and typically lasts half or one day.

> *In future, we will not really be able to know everything there is to know. Quite often, we will have to make judgements based on the information we are given and be able to account for the judgements that we have made and the actions we have taken. This will be a fundamental skill for the workforce of the 21st Century.*

> *Its value lies primarily in that it is applicable to every other academic school subject I have come across, … many that pupils will not encounter until university or later life.*

> *The pupils have poor analytical skills and believe most conspiracy theories and media headlines shown to them. They are reluctant to analyse what they read on the web in particular. This should be a growing concern and CT combats this to some extent.*

Of the teachers who reported that they thought that Critical Thinking was 'not at all' or 'slightly' valuable (n=2 and n=8 respectively), some quotes are included.

> *Only done really as a means to a qualification and to help Oxbridge students with tests and interviews. Might be more useful if done at greater length for other reasons.*

> *The narrow subject specific definitions used limit its usefulness in other subjects.*

Some of the comments on why students find it highly or moderately valuable include:

> *Those attending see it not just as an extra AS but as an opportunity to discuss, debate and generally have their minds expanded by exposure to materials which challenge and focus their thinking skills without distraction by masses of subject-specific rote-learning.*

> *Where else do you get to argue openly with a teacher?!*

> *They vary. Some think it's crap. Others really see the point of questioning the propaganda and the spin.*

> *They no longer blindly accept what they are told. It sparks discussion and debate.*

> *They transfer the skills to other subjects and are aware that in combination with subject knowledge they can develop a powerful tool for analysing discourse.*

Common themes reported were that students simply enjoy CT (n=24) or believe they have beneficially transferred the skills into other subjects (n=20). Another common theme (as mentioned elsewhere in this article) was that for students, the value of a subject lies within its currency for a university place, though there were divergent views on whether CT added to their application or not. Consider the juxtaposition of the following two quotes:

> *Feedback from students has been positive. They understand the value placed on Critical Thinking by HE organisations and many have commented on how it has strengthened their understanding in other subjects.*

> *They see it as a good A level to have until some unis unhelpfully say they will not consider it.*

Some of the comments accompanying less positive student ratings are included below. Common themes once again include identifying the value of the subject with UCAS points, perceived difficulty, and an instrumental approach to learning. A few responses alluded to the nature of the assessment limiting the students' ability to engage with the subject.

> *The vast majority of pupils regard "usefulness" as meaning "is it useful to the UCAS process?" and the overwhelming feedback… is that universities are not interested in it.*

> *They tend to think in an instrumental way and not think about learning as an activity that has intrinsic value.*

Another theme was that low valuation on behalf of the students was (partly) a result of the limited timetabling:

> *It only takes a look at the timetable for students to make up their mind about how valuable the subject is in comparison with other subjects.*

The most negative comments concerned those students for whom CT was mandatory.

> *They see it as a forced option and hate it. Lessons not particularly stimulating as a one-term rush inevitably has to be focussed on exam-practice.*

Finally, a quote that encapsulates how a nexus of factors can contribute to a negative valuation of the subject:

> *I fear that their utilitarian attitude to exams/courses rather holds them back. They are so highly examined – rather trapped in the system to the extent that they can not always see the point of doing anything that 'doesn't count' on August 16th. Also the disappointing results have been a real blow. They ask what is the point of doing a hard subject. They only want As and Bs and see anything else as an insult.*

Teachers were asked about their enjoyment of teaching CT (84% were positive) and 186 respondents provided additional comments regarding their enjoyment. Many teachers' enjoyment of Critical Thinking derived from their view that it is 'new', 'different' or 'fresh' (n=28):

> *It's been a "shot in the arm" for a teacher who needed a new stimulus.*

> *It is refreshing to be able to encourage children to actually use their brains rather than just worry about memorising information and 'getting the right answer'. It is exciting to see them grow in confidence and skills.*

… or simply just fun/enjoyable (n=33):

> *It allows me to indulge myself in the "Dead Poets Society" aspect of teaching which I particularly enjoy.*

In particular, (and this was the most common theme) teachers tended to describe a greater freedom or creativity in choosing materials to teach Critical Thinking because of its skills-based nature (n=40).

> *I enjoy the freedom from the drill of a body of knowledge but enjoy the discipline that the skills provide. It seems to me that this subject develops the skills that have been squeezed out of other subjects by the national curriculum.*

However, not all teachers responded so positively to this aspect of teaching the discipline. Two respondents found the lack of 'factual' content a drawback, for example:

> *I enjoy teaching it but find it very challenging and it definitely moves me out of my comfort zone. Not having specific content or definitive answers takes some getting used to.*

Frequently, teachers referred to the benefits for the students (such as the transferability of skills into other subject domains), and the perceived 'worthwhile' nature of the subject (n=23). A number of teachers highlighted the enjoyment they derived from challenging, stretching and encouraging students' thinking (n=23).

> *It gives me the opportunity to challenge the brightest students and to develop their intellect far more than is possible at KS4.*
>
> *I like encouraging thinking – education should, as Hemingway put it, "make you a good crap detector"...*

A number of teachers (n=21) commented that their enjoyment stemmed from being stimulated and challenged themselves (in a positive way):

> *... Having taught for 23 years it is a new challenge to me (and I do find it challenging at times) and I am learning a lot myself through teaching it – that is very much part of the enjoyment.*

As with the last quote, a number of teachers believed that through teaching Critical Thinking they were developing professionally: they were upskilling in terms of their own thinking skills as well as professional situations (n=15). This is potentially an important and unanticipated collateral benefit of teaching Critical Thinking.

> *Teaching this subject has altered the way I think. I find myself using the skills not only in the classroom but also in meetings and other aspects of my life.*
>
> *Made me be more rigorous in my own thoughts.*

Additionally, on a professional front, teachers reported enjoyment of adopting different teaching styles/pedagogical approaches (n=18) – more student-centred, interactive and less didactic. This finding echoes both Sternberg's views of CT pedagogy (see earlier) as well as the findings of Baumfield and Oberski (1998) that (broader) thinking skills approaches in classrooms were popular with teachers because they foster changing patterns of interaction in the classroom. Teachers commented that adopting this different role in the classroom meant that they learnt a lot from the students and that they welcomed the greater unpredictability and 'uncertainty' in lessons.

> *It's a subject you can 'discuss'; it requires little didactic teaching which is good.*
>
> *I really like challenging myself in the teaching and sometimes I do not know the answers and work them out with the students. I find that very powerful as a teaching tool and a model for learning.*
>
> *The chemistry between students and students-tutor brings a levelling as ideas and argument can arise from any of many sources. The tutor as 'facilitator' is attractive and (when it works!) is very fulfilling.*
>
> *More unanswerable questions are raised than in other subjects and there is a real opportunity to challenge and explore each other's points of view.*
>
> *Every lesson is different – I am always surprised or stimulated by student responses.*

However, taking such a role and operating in a more 'uncertain' classroom was not comfortable for all respondents (n=2):

> *It's not like teaching other subjects where you can hide being wrong or not knowing: students lose faith in your ability to teach; this has implications for the senior role I play in college.*

This interesting comment resonates with Blagg's observation (1991) that the feeling of being 'deskilled' is more of a threat to an experienced teacher than it is for a novice.

A number of factors were mentioned for tempering or, in a few cases, eliminating enjoyment of teaching the subject. One common 'negative' theme was encountering problems with the materials – either accessing materials or that the materials available were considered too 'dry' (n=12):

> *Need far more resources than are at present available.*
>
> *[Students] only show an interest when I provide material I've adapted – which is very time consuming to produce.*

Again, there were also some issues with timetabling limitations (n=11), or, as a teacher's second or third subject, prioritisation (n=3):

> *With very limited time available I have not been able to do many of the activities I would have liked to do.*
>
> *I would LOVE to teach Critical Thinking properly but I am not given the time on the timetable, the teacher-resources, or the support I require in school either to teach my own classes properly or to co-ordinate the delivery of it school-wide.*
>
> *I do enjoy teaching CT but as my other subjects (History and Politics GCSE and A Level) take priority... Therefore I feel the students do not always get the best deal in CT lessons.*
>
> *I am not trained to teach it. It is not my priority. Students attend poorly and show little interest.*

Finally, the last fixed-choice question in this section asked whether respondents believed that Critical Thinking skills can benefit students in their other AS exams.

The overwhelming majority of respondents answered yes (see Figure 6).

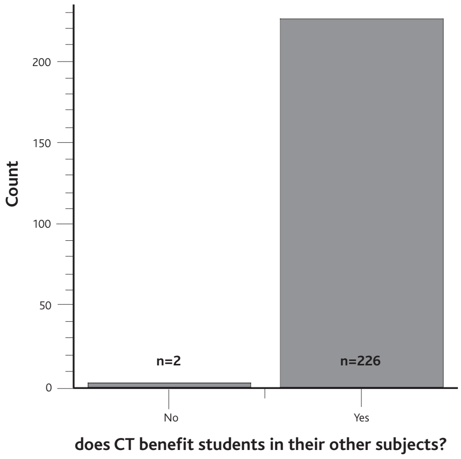


Figure 6: frequency bar chart of respondents' views on whether CT benefits students in their other AS subjects

Accompanying comments (n=186) were mainly positive (or, indeed, very positive), with only a few showing some equivocation. Comments tend to highlight subjects that can particularly benefit and/or the skills which are particularly transferable, or in some cases describe how other staff or students themselves have ascribed increased performance in other subjects to Critical Thinking.

> *It can but I am not sure it does. The Heads in both schools where I teach Critical Thinking believe it improves A Level results. I don't have the data.*

> *Many subjects call for reasoned arguments. What better way to prepare them?*
>
> *Making cross curricular links is highly useful. It also encourages them to think more broadly about their work and how to approach it.*
>
> *… the majority [of students] find it quite useful and they now write better essays or think more logically. One said 'It has changed my whole way of thinking'.*
>
> *Complements analytical requirement in many subjects… Many of our "most-improved" students in year 13 took CT in their year 12 perhaps due to developing transferable skills.*

Many of these assertions indicate that Critical Thinking is, or at least, is believed to be a powerful educational force.

Discussion

Because the respondents were self-selecting, it is more than likely that they do not represent the full range of teachers of Critical Thinking. Many of the respondents identified that they were the co-ordinator or the sole teacher of Critical Thinking in the centre. It seems likely that the ordinary 'foot soldier' is under-represented in the sample. This possibly may explain the disparity between Vidal Rodeiro's findings that Critical Thinking is often poorly valued by students, and the findings from this research. Certainly, in this study, the centres where Critical Thinking was mandatory reported significantly lower levels of student motivation, attendance and enjoyment.

The responses to the questionnaire identified a series of obstacles and challenges which teachers of Critical Thinking have been faced with, many of which interact together. A frequent theme was the value placed upon Critical Thinking (c.f. Vidal Rodeiro, 2007). Interestingly, while Richardson (2008) notes that, for Citizenship, formal summative assessment was perceived as being needed in order to 'credentialise' a new subject, our findings make it clear that formal summative assessment is far from sufficient. In order for the subject to acquire the same 'credentials' as other AS qualifications, it is not enough that there is an exam, and that the subject/exam is perceived to be difficult. It needs endorsement from universities' admissions policies, as well as centres themselves offering the course in a fully resourced and supported manner. Evidently though, some teachers can and have overcome some of these obstacles by promoting the perceived intrinsic value of the subject, and many appear to be passionate advocates of the value of the discipline both in terms of its life skills and transferability to other academic subjects. This report shows that many of the respondents have been (and have had to be) very resourceful in terms of training themselves in this new subject. They have responded positively to the greater freedom in lessons and have altered their teaching styles in order to deliver it (c.f. Baumfield and Oberski, 1998; also noted in Blagg, 1991). There is some evidence that by teaching Critical Thinking, teachers themselves were developing professionally: they were up-skilling in terms of their own thinking skills, in terms of using greater analysis and evaluation skills in other subject lessons as well as in other professional situations. This is a potentially an important and unanticipated collateral benefit of teaching Critical Thinking and possibly warrants further investigation.

The challenge to any new subject lies in it finding its 'niche' within the curriculum. As for AS Critical Thinking, it is in a paradoxical position. It is like other AS levels in that it leads to the 'same' qualification – an AS level. And yet, it does not have true parity with them, first because of the nature of its subject matter (it has a higher focus upon skills and lower focus upon content), secondly because it is often delivered in much less contact time, and thirdly, given its current status with universities, many students would probably want to avoid taking Critical Thinking as one of their main AS levels. However, if it were not an AS level, it is probably true to say that many fewer students would have had the opportunity to study Critical Thinking.

We conclude with some speculations about what the future might hold for Critical Thinking in schools. As time goes on, teacher experience and expertise in the subject will accumulate, and a greater range of resources will be available. This should have a positive impact upon teaching and learning. However, this can only happen if the strategies in schools permit it. Thus, where schools 'drop teachers in the deep end' at the beginning of the school year (as several respondents reported), do not support teachers in terms of funding resources, sufficient timetabling or training days (again, reported in this study), this vital accumulation of expertise is prevented from happening.

Perhaps the key matter for the future success of Critical Thinking AS level is for it to gain greater acceptance with universities. Currently, its acceptance as part of a 'main offer' is patchy. As reported above, this is a significant source of frustration for teachers who do see its value, but who have to deal with students' consequential low motivation. If universities were to more widely acknowledge its value and endorse its status, the future for Critical Thinking would be much more secure.

References

Abrami, P.C., Bernard, R.M., Borokhovski, E., Wade, A., Surkes, M.A., Tamim, R. & Zhang, D. (2008). Instructional Interventions Affecting Critical Thinking Skills and Dispositions: A Stage 1 Meta-Analysis. *Review of Educational Research* **78**, 4, 1102–1134.

Baumfield, V.M. & Oberski, I.M. (1998). What do Teachers Think about Thinking Skills? *Quality Assurance in Education*, **6**, 1, 44–51.

Black, B. (2008). *Critical Thinking – a definition and taxonomy for Cambridge Assessment: supporting validity arguments about Critical Thinking assessments administered by Cambridge Assessment*. Paper presented at 34th International Association of Educational Assessment Annual Conference, 9th September 2008, Cambridge.

Blagg, N. (1991). *Can we teach intelligence? A comprehensive evaluation of Feuerstein's Instrumental Enrichment Program*. London: Laurence Erlbaum.

Bloom, B.S. (Ed.), Engelhart, M.D., Furst, E.J., Hill, W.H. & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. New York: David McKay.

Kerr, D. & Cleaver, E. (2004). *Citizenship Education Longitudinal Study: Literature Review – Citizenship Education One Year On – What does it mean? Emerging definitions and approaches in the first year of the National Curriculum Citizenship in England* (Vol. 532). Nottingham: DfES.

Richardson, M. (2008). *Assessing citizenship: measuring the unmeasurable?* Paper presented at the Fourth Biennial EARLI/Northumbria Assessment Conference, Berlin, Germany, August 2008.

Sternberg, R.J. (1987). Teaching critical thinking: eight ways to fail before you begin. *Phi Delta Kappa*, **68**, 246–9.

Vidal Rodeiro, C.L. (2007). *A-level subject choice in England: patterns of uptake and factors affecting subject preferences. Research Report*. Cambridge: Cambridge Assessment.

Response to Cambridge Assessment's seminar on Critical Thinking, February 2010

Joe Chislett

Joe Chislett is a senior examiner in Critical Thinking and a teacher at Westminster Kingsway College

Cambridge Assessment recently organised a seminar, hosted at the British Academy, on the role and value of Critical Thinking and its impact upon driving attainment. Many interesting questions and issues were raised, one of the most interesting being whether or not Critical Thinking could, or indeed should, be 'embedded' into other subjects, rather than taught and assessed as a standalone subject in its own right.

As someone who has taught Critical Thinking for ten years alongside A levels in English and Philosophy, and who has been involved in the Cambridge Assessment definition and taxonomy work, I do clearly have an allegiance to the subject. Nevertheless, it is my conclusion that many of the skills of Critical Thinking cannot be effectively taught by just embedding them in other subjects; and the question we must ask is – to what extent do we value these skills? To start with, I will try to clarify an area of confusion that I think distorted the debate, and that I believe has influenced the arguments of those who feel Critical Thinking can be successfully embedded in the 'proper' study of other 'proper' subjects. I would like to emphasise that, as a Critical Thinking enthusiast, I am not being protective about my subject area. As I hope this article will show, I value deeply the skills Critical Thinking teaches, and if they can be delivered through other subjects, that is excellent: I do not wish to hang onto them!

The arguments against the teaching of Critical Thinking as a standalone subject rested on two main premises. One was that thinking and reasoning needed a context: something to think and reason about. This is obviously true; it is a bit like saying you cannot practise passing a football around without a football. (Although it is worth noting that you can practise passing a football without engaging in a football match.) The more important objection was that the skills Critical Thinking teaches are those that are, or at least should be, acquired through the study of other subjects. To understand the force of this objection, we need to make a distinction between two senses of 'Critical Thinking': as a set of skills, and as a set of dispositions.

There are two objectives in teaching Critical Thinking. One is dispositional: to encourage an open-minded, critical, independent, healthily sceptical and questioning outlook; in short, to encourage people to think. The other is to encourage people to think *well*.

If we mean by Critical Thinking *just* the dispositional approach, then of course this can be embedded. It is, in the absence of specific Critical Thinking skills, really no more than an approach to study, to the way subjects are taught and assessed. As a teacher, I regularly come across students who strongly exhibit this approach to learning. It is true that they are not the norm; at my college, I usually encounter no more than one or two every year, and it would be nice if there were more of them. However, it is quite common for the students with this dispositional outlook, while they are naturally inclined towards thinking and reasoning for themselves, to think and reason badly. Wherever possible, I encourage them to take Critical Thinking. There was one student I taught for English last year. She was extremely – fiercely – independent-minded; and yet her arguments and thinking were often horribly flawed. Occasionally I would try to challenge her; to point out her reasoning errors; but it was generally not possible to do so. This is not simply because of the constrictions of the subject; it is simply not something you can 'tack on'. I would have needed to devote several hours to the concepts of reasoning, argument, inference and logic. I would, in short, have needed to stop teaching English and start teaching Critical Thinking.

The arguments I have heard in support of making Critical Thinking embedded or implicit in other subjects seem to me to have conflated the notion of critical thinking as a disposition with critical thinking as a set of specific skills. There is also the assumption that, since all academic subjects entail thinking skills (along the lines outlined in the Cambridge Assessment taxonomy), this means that pursuing these subjects will teach students how to perform these skills well. There are two reasons why this assumption is mistaken.

First, it is becoming increasingly evident through the study of the human mind and its reasoning patterns that we all as humans, even high level academics, have innate tendencies to reason poorly. One of the first things to convey in teaching Critical Thinking is that thinking and reasoning effectively is *difficult*. (The fact is that year on year, the same students at my college, with the same teaching time, tend to do slightly worse in Critical Thinking than their other subjects. They find it *more difficult*: they admit this.) For most people, if not everyone, correct forms of reasoning are often counter-intuitive. Almost everyone has a tendency to (what is known in logic) 'affirm the consequent', and a weak tendency also to 'deny the antecedent'. These are both invalid forms of reasoning; but for psychological reasons are, to the untrained eye, utterly compelling. To correct these and many other kinds of inbuilt reasoning errors we make takes time and specialist input. It is not just encouraging people to 'think for themselves'; this will only lead to their own bad patterns of reasoning becoming more deeply entrenched. People can and do reason poorly despite achieving great success in their own specialist fields. It sometimes only takes a little explicit input on forms of reasoning before students are able to see and explain the flaws and errors in reasoning made by, presumably, well-educated individuals, such as academics, scientists, politicians and journalists. Explicit training really helps. When I used to defend my choice of Philosophy as a degree, one of the strongest points in its favour was that it helped me to think clearly, logically and analytically (or at least more so than I would have done otherwise). And yet, the challenges of Critical Thinking AS level have helped me significantly *beyond* my degree. It has helped me also to understand, and to teach, my other subjects better.

Secondly, the importance of having explicit training in thinking, reasoning, or logic is that we ought to value thinking as an end in itself. We should value thinking, value our reason and rationality, as an excellence in itself; not as something that is simply the by-product of a particular academic discipline. On it depends our own autonomy. Yes, it does underlie specialist subjects – so it will (and does) enhance what is done in each of those. But more importantly, it underlies what it means to be human.

A tricky task for teachers: assessing pre-university students' research reports

Irenka Suto Research Division **and Stuart Shaw** CIE Research

Introduction

In the UK and internationally, many students preparing for university are given the challenge of conducting independent research and writing up a report of around 4000 or 5000 words. Such research activities provide students with opportunities to investigate a specialist area of study in greater depth, to cross boundaries with an inter-disciplinary enquiry, or to explore a novel non-school subject such as archaeology, cosmology or anthropology. We theorise that, as is the case in higher education (Brown *et al*. 1997), independent research encourages intellectual curiosity whilst enabling students to develop skills in practical and analytical research, higher order thinking, interpretation and time management. When applying to university, students can use their reports to demonstrate motivation for their intended course of study and to differentiate themselves from competing applicants.

In the wake of the recommendations of the Tomlinson Report (2004) on the shape of 16–19 qualifications in England, The Sixth Form College, Farnborough, developed a systematic approach to encouraging its students to conduct independent research. Since 2006, students have been carrying out extended projects during their holidays or alongside their other courses, generating formally-structured reports. The reports are assessed formatively through detailed written comments to the students by their teachers, rather than assessed summatively by issuing a mark. This has generated a considerable body of student evidence within the college.

At other schools, students conduct projects which constitute or contribute to a formal qualification, and which are therefore assessed summatively. For some of these qualifications, the students' research reports are assessed by their own teachers. The teachers' marks are then moderated by professional examiners who are employed by the examination board administering the qualification. The Cambridge Pre-U Independent Research Report, administered by Cambridge International Education, utilises this assessment approach, as do the extended projects administered by the AQA, OCR, and Edexcel examination boards. Extended projects can be used to obtain a stand-alone qualification. Alternatively they can contribute to a 14–19 Diploma in England or the Welsh Baccalaureate qualification in Wales. For other qualifications, such as the International Baccalaureate, students' research is marked exclusively by external examiners.

The assessment of research reports poses several challenges, including those which arise when assessment schemes are designed to reward generic research skills rather than particular subject knowledge. Assessors may lack detailed understanding or marking experience of the research topics explored by some students. However, it is unclear whether subject knowledge facilitates or hinders marking. For example, familiarity with particular terminology or technical language may aid interpretation of what the student has written. Alternatively it may obscure the assessor's perception of generic skills, especially if they have been mis-applied by the student.

In this study, we explored the feasibility of applying a single mark scheme to research reports covering diverse topics in order to reward generic research skills. Our aim was to investigate the reliability with which teachers can mark diverse research reports, using four different generic assessment objectives. We also investigated teachers' views in applying generic mark schemes, particularly when marking reports on unfamiliar topics.

The Cambridge Pre-U Independent Research Report (IRR)

The study was conducted as part of a wider on-going research programme supporting the Cambridge Pre-U, a new type of qualification for 16–19-year-olds which is designed to equip students with the skills required to make a success of their university studies. The first cohort of Cambridge Pre-U students will be completing their courses in the summer of 2010. Typical Cambridge Pre-U students study three Principal Subjects over a two-year period (or alternatively, a combination of Principal Subjects and A levels). In addition to this, to obtain the Cambridge Pre-U Diploma, they must complete the Cambridge Pre-U's course in Global Perspectives and Independent Research (GPR).

GPR is known as the core of the Cambridge Pre-U Diploma but also constitutes a stand-alone qualification with a UCAS tariff equivalent to an A level. It comprises two components: the Global Perspectives course (GP), and the Independent Research Report (IRR) which may be up to 5000 words long. The GP and IRR have been designed to provide students with coherence, depth and breadth, through encouraging focused personal exploration and increased depth of study. They expand creative, critical and responsible awareness through the tackling of different perspectives on global issues. Assessment of the IRR focuses on the student's abilities in a range of areas. These include: designing, planning and managing a research project, collecting and analysing information, evaluating and making reasoned judgements, communicating findings and conclusions, and uniquely, intellectual challenge. The present study explores the practical application of four different generic assessment objectives which comprise a substantial proportion of the mark scheme that will be used to mark the IRR this summer.

Participants

Fifteen teachers (10 men and 5 women) participated as markers in the study. They were recruited by e-mail from nine different schools in England whose 16–19 year-old students were either currently working on independent projects or planning to do so in the near future. The teachers had a wide range of subject backgrounds and teaching and examining experiences.

The teachers' experimental marking was led by a highly experienced examiner: the Chief Examiner (CE) for Cambridge Pre-U's GPR course, who also undertook this role in the study.

Project reports

The study was conducted prior to the completion of any Cambridge Pre-U IRRs by Cambridge Pre-U students. We therefore explored the marking of project reports produced by students of The Sixth Form College, Farnborough, UK. Like IRRs, the projects could be on any topic of interest to students, the reports had an approximate word limit of 5000 words. However, as the projects did not contribute to any qualification, the students had not written the reports with any particular assessment objectives or marking criteria in mind.

The college provided the researchers with copies of 346 project reports (68 from 2006, 135 from 2007, and 143 from 2008). At a two-day meeting, the researchers and CE jointly reviewed the reports and selected a sample of 20, stratified by subject area. From these 20 reports, a sub-sample of 5 was selected for use by participating teachers as a practice sample. Full details of the report selection process are given in the appendix.

The CE determined a fixed marking order for the 5 reports in the practice sample, which were numbered accordingly. The remaining 15 reports comprised the main sub-sample. The researchers determined a random marking order for these reports and numbered them accordingly. The report titles are shown in Table 1.

Mark scheme

An experimental version of a mark scheme was used in the study which was derived from that for the Cambridge Pre-U IRR. The original IRR mark scheme is divided into five Assessment Objectives (AOs, see Table 2), enabling assessment of each of the five AOs at three different levels. Since for AO1, students are required to "design, plan, manage and conduct own research project using techniques and methods appropriate to the subject discipline", AO1 can only be assessed in the context of the classroom, by students' own teachers. As the study's teachers were to mark the work of students they had not taught, AO1 was omitted in the experimental mark scheme. Similarly, part of AO4 relates to a student's negotiation with his/her tutor; as it could not be used in this study, it was removed from the experimental mark scheme.

Table 1: Titles of project reports used in the study

Sub-sample	*Report number*	*Project report title*	*Broad subject area*
Practice	01	Can we trust Quantum Theory over Electromagnetic Wave Theory of Light?	Physics
	02	Would the British economy have been as successful without the transatlantic slave trade?	Economics
	03	Is prison the best sentence for paedophiles, or do alternatives offer a safer and more effective rehabilitation option?	Criminology
	04	Addiction – nature or nurture?	Psychology
	05	Polya's heuristics: are they applicable in a broader context?	Mathematics
Main	06	How effectively has Ghana dealt with the problem of malaria?	Geography
	07	An exploration into the role of metaphor in economics	English
	08	Is prescribed medication the most effective way to treat Attention Deficit Hyperactivity Disorder?	Biology
	09	Does the French language need protecting, and if so is enough being done to protect it?	French
	10	Is it right to chemically alter the behaviour of children through the use of drugs such as Ritalin?	Biomedical ethics
	11	Has Pina Bausch revolutionised ballet with her controversial 'Tanztheater'?	Drama
	12	Hydrogen fuel: can hydrogen replace gasoline?	Chemistry
	13	Should the UK join the Euro?	Politics
	14	Could an artificial intelligence be an ideal ruler?	Philosophy
	15	Can a murderer's behaviour be reduced down to biological or environmental factors, or is it a combination of both?	Psychology
	16	Is communism viable today?	Politics
	17	Is punk rock art?	Art
	18	Should permission be given to remove the treatment of patients in a persistent vegetative state?	Biomedical ethics
	19	What philosophical problems arise with Chomsky's account of language acquisition?	Linguistics
	20	To what extent does music have a beneficial effect on brain activity?	Music

Table 2: Assessment objectives and marks in original mark scheme

Assessment Objective	*Domain*
AO1	Knowledge and understanding of the research process
AO2	Analysis
AO3	Evaluation
AO4	Communication
AO5	Intellectual challenge

Procedure

The experimental procedure comprised the following stages:

1. The Chief Examiner (CE) marked all 20 reports, thereby generating a 'correct' mark for each one.
2. Each teacher was posted the sample of 20 numbered reports, together with the mark scheme, practical instructions about the study from the researchers, and detailed written guidance on marking from the CE. A marking grid was also provided, to be used to record marks and notes.
3. Each teacher began by familiarising him/herself with the mark scheme and reading the CE's guidance on marking.
4. Each teacher marked the practice sub-sample (N = 5) in numerical order, recording his/her level followed by his/her mark and notes in the marking grid. Teachers were welcome to annotate the reports.
5. Each teacher contacted the CE, who provided personalised telephone feedback on his/her marking of the practice sample. Teachers were asked to record the CE's marks and feedback in their marking grids. The CE also kept records of the teachers' marks and the feedback given.
6. After receiving telephone feedback, each teacher marked the main sample (N = 15) in numerical order. The teachers were asked to try to apply the CE's advice wherever possible. For each report, they recorded their marks and notes for each assessment objective in the marking grid. Again, the teachers could annotate the reports if they wished.
7. After completing the marking, each teacher filled in a questionnaire about his/her marking experiences.
8. All documents were returned to the researchers.

Analysis and findings

All 15 teachers marked all 20 reports in the study. However, one teacher had to withdraw from the study for personal reasons prior to completing the post-marking questionnaire. Analyses were conducted on the marking of the main sub-sample and the questionnaire data using SPSS Version 15.01 and FACETS Version 3.6 software.

Correlation of marks

Indices of inter-rater reliability among all participants (i.e. the 15 teachers and the CE) were calculated for each of the four Assessment Objectives (AO2–5) and for the total score using a procedure described by Hatch and Lazaraton (1991, p.533). This entailed generating a Pearson correlation matrix for all participants for each AO. A Fisher Z transformation was then applied to the correlations, to transform the correlations to a Normal distribution and to correct the distortion inherent in using the Pearson for ordinal data. The mean correlation among participants could then be calculated. Subsequently, the derived mean of the transformed correlation coefficients, r_{ab} was substituted into the formula:

$$r_{tt} = \frac{n.r_{ab}}{l+(n-l)r_{ab}}$$

where r_{tt} stands for the reliability of all the participants' ratings, n is the number of participants, and r_{ab} is the average correlation among participants. Finally, r_{tt} was transformed back to a Pearson's correlation coefficient.

Table 3: Inter-rater marking reliabilities (among all participants)

	Number of marks available	*Pearson's correlation coefficient*
AO2	18	0.71
AO3	18	0.72
AO4	9	0.71
AO5	6	0.73
Total score	**51**	**0.72**

Table 3 presents the mean correlations for each AO and for the total score.

These reliability figures compare favourably with those estimated and reported elsewhere. For example, Shaw (2008) quotes inter-rater reliability indices of 0.78 using the same statistical approach. In another, similar study investigating marking reliability of essay questions from the higher tier of GCSE English Literature, Johnson, Nádas and Bell (2009) also report reliabilities of a comparable magnitude. However, these studies both focus on medium length constructed responses which are considerably shorter than the 5000-word reports used in the present study. The focus of a study by Laming (1990) offers a closer comparison. Laming's investigation was designed to estimate reliability between pairs of examiners marking a university examination comprising a number of extended essay-type answers. Laming found that the correlation between the marks independently awarded by pairs of examiners varied between 0.13 and 0.72. Given the participants' lack of familiarity with the present study's experimental mark scheme, the reliability figures calculated here are encouraging.

These findings were corroborated by a statistical check employing multi-faceted Rasch analysis. In the context of inter-rater reliability, FACETS models participants as 'independent experts'. Although FACETS does not estimate inter-rater reliability directly, it routinely generates observed and expected agreement percentages. Adapting Cohen's Kappa agreement statistic enables the estimation of a Rasch-based Kappa coefficient. Under Rasch-model conditions ideally this should be close to 0, indicating that inter-rater reliability is within the acceptable range.

The Rasch-Cohen's Kappa is calculated as:

$$\frac{\text{(Observed agreement \% – Expected agreement \%}}{\text{(100 – Expected agreement \%)}}$$

Values of Rasch-Cohen's Kappa for each AO are presented in Table 4.

Table 4: Values of Rasch-Cohen's Kappa for AOs

Assessment Objective	*Rasch-Cohen's Kappa*
AO2	0.0088
AO3	0.0212
AO4	0.0038
AO5	0.0160

These values are close enough to 0 to support the previous findings of high reliability for report marking.

In order to explore participant agreement further, FACETS was used to provide two measures of 'fit' (or consistency): the 'infit' and the 'outfit' values.[1] There are different views on what fit index is actually acceptable. McNamara (1996) suggests that the usual limits of acceptability are the mean ± 0.3 (so anything between 0.7 and 1.3 is acceptable). According to Lunz and Wright (1997, p.83) "Because the interpretation of fit is situationally dependent, there are no fixed levels for fit statistic acceptance or rejection." They go on to use a level of ± 0.5 in their studies. Wright and Linacre (1994, p.370) suggest figures ranging from 0.4 for 1.7 depending on the type of assessment under investigation: fit statistics of 1.7 or greater indicate too much unpredictability in raters' marks, while fit statistics of 0.4 or less indicate overfit or not enough variability in raters' marks. The infit and outfit values for the CE and 15 teachers were calculated for each AO. Overall, given the above guidance on levels of fit, they indicated a generally well-fitting Rasch model.

When considered together with the descriptive statistics and estimations of inter-rater reliability, the Rasch findings reveal a good degree of agreement among participants on each of the four AOs.

Relative marking severity and variation

For each report, the CE's marks were deemed to be correct and therefore the 'gold standard'; they were used as the comparators against which all teachers' marks were compared. This analysis explored marking agreement with a consideration of two descriptive statistics:

- *marking mean* – a measure of relative severity of the marking.
- *standard deviation* – a measure of the range of marks used. The larger the standard deviation, the wider the range of marks awarded.

Table 5 summarises the mean total marks given by each participant to the 15 reports. On average, the CE's total marks are lower than those awarded by the teachers and cover a narrower range. ANOVA revealed a significant difference among the participants (F = 2.36, d.f. = 15, 224, p < 0.05); however, deeper investigation with post-hoc tests (Bonferroni and Tukey) indicated that only one teacher (G) marked significantly more severely than the others.

An analysis of the marks awarded on individual assessment objectives was also conducted. Both AO2 (Analysis) and AO3 (Evaluation) employ a mark range of 1–18 marks across three levels. The mean marks in Table 6 and Table 7 indicate the relative severities of the 15 teachers and CE on these two AOs.

Table 5: Descriptive statistics for the total marks given by participants

Teacher	*Main subject(s) taught*	*Mean mark*	*Standard deviation*
CE	History	26.93	9.05
A	Critical thinking	31.60	9.98
B	History, politics, business studies	25.07	11.95
C	Law, politics, psychology	28.20	9.44
D	History	28.67	10.55
E	Religious studies, philosophy	33.27	9.07
F	Philosophy, ethics, religious studies	31.07	8.96
G	Physics, astronomy	22.93	8.36
H	English, media studies	31.53	9.58
I	English	29.07	8.48
J	Maths	34.53	10.72
K	Politics, history, critical thinking	25.73	9.84
L	Biology, chemistry	23.27	10.77
M	Theory of knowledge, classical civilisation	33.40	10.24
N	English, critical thinking	35.67	7.58
O	Chemistry	28.60	11.97

Table 6: AO2 Descriptive statistics

Teacher	*Mean mark*	*Standard deviation*	*N reports marked*
CE	8.20	3.28	15
A	10.33	4.06	15
B	8.53	4.45	15
C	9.60	3.64	15
D	10.33	3.83	15
E	11.40	3.58	15
F	10.80	3.19	15
G	6.13	2.61	15
H	10.13	3.66	15
I	8.80	3.28	15
J	11.47	3.81	15
K	7.73	3.79	15
L	7.60	3.98	15
M	10.93	3.90	15
N	12.67	2.82	15
O	9.07	4.62	15

Table 7: AO3 Descriptive statistics

Teacher	*Mean mark*	*Standard deviation*	*N reports marked*
CE	8.93	3.99	15
A	10.73	4.64	15
B	8.20	4.48	15
C	10.00	3.44	15
D	8.47	4.73	15
E	11.53	3.04	15
F	10.53	3.36	15
G	8.40	3.44	15
H	10.80	3.99	15
I	10.40	3.89	15
J	11.80	3.84	15
K	9.53	4.22	15
L	8.27	4.40	15
M	11.40	3.72	15
N	11.93	3.17	15
O	9.60	4.78	15

For AO2, the mean marks ranged from 6.13 to 12.67. ANOVA revealed significant differences among the participants (F = 3.24, d.f. = 15, 224, p < 0.05); post-hoc tests indicated that Teachers G, K and L marked significantly differently from the others. The table shows a spread in standard deviation of nearly 2 marks when assessing AO2.

Whilst there were differences in severity among teachers in the marks awarded for AO3, these were less marked than for AO2 and not statistically significant (F = 1.61, d.f. = 15, 224, p >.05), that is, the participants generally behaved as a homogeneous group. Although AO3 and AO2 are equally weighted, the tables reveal a greater spread of marks for AO3, suggesting that AO3 is discriminating among reports more effectively.

In general, the CE tended to mark more harshly on both AO2 and AO3 than the teachers do, although this tendency is less pronounced on AO3 and over a slightly narrower range on AO2.

AO4 (Communication) is assessed against a 9 mark scale. As Table 8 shows, the trend towards CE severity (apparent for AO2 and AO3) is reversed in the case of AO4 where teachers tended to be slightly more severe than the CE.

AO5 is assessed against a 1– 6 mark scale, which is the shortest scale. Evidence from the marks (Table 9) suggests that, on average, the CE marked more harshly on AO5, and over a slightly wider range, than the

Table 8: AO4 Descriptive statistics

Teacher	*Mean mark*	*Standard deviation*	*N reports marked*
CE	6.73	1.71	15
A	6.47	1.73	15
B	5.40	2.10	15
C	5.80	1.42	15
D	6.47	1.41	15
E	6.33	1.72	15
F	5.73	1.67	15
G	5.53	1.85	15
H	6.60	1.24	15
I	6.60	1.76	15
J	6.93	1.94	15
K	5.73	1.49	15
L	5.07	2.12	15
M	6.87	1.68	15
N	7.00	1.31	15
O	5.87	2.23	15

Table 9: AO5 Descriptive statistics

Teacher	*Mean mark*	*Standard deviation*	*N reports marked*
CE	3.07	1.39	15
A	4.07	1.28	15
B	3.00	1.51	15
C	2.80	1.26	15
D	3.40	1.72	15
E	4.00	1.25	15
F	4.00	1.13	15
G	2.87	1.41	15
H	4.13	1.13	15
I	3.27	1.03	15
J	4.33	1.50	15
K	2.80	0.94	15
L	2.47	1.13	15
M	4.20	1.37	15
N	4.07	0.88	15
O	3.80	1.86	15

[1] The infit is the weighted mean-squared residual (the difference between actual marks and marks predicted by the Rasch model) which is sensitive to unexpected responses near the point where decisions are being made, while the outfit is the unweighted mean-squared residual and is sensitive to extreme scores. For ease of interpretation, the two sets of fit statistics are expressed either as a mean square fit statistic or as a standardised fit statistic, usually a *z* or *t* distribution.

teachers. As with AO2, ANOVA revealed significant differences among the participants (F = 3.28, d.f. = 15, 224, p <.05); post-hoc tests indicated that Teachers J, L and M marked significantly differently from others.

The scatter diagram in Figure 1 shows the relationship between the mean of the teachers' total marks and the CE's (gold standard) total marks. If the two marking approaches were to yield identical marks, then the points on a scatter diagram would all lie on a *line of identity*, shown with a dotted line in Figure 1. It can be seen that ten points lie above the identity line, indicating frequent marking leniency relative to the CE reports. Very few points lie below the identity line, indicating that marking severity relative to the CE was much rarer.

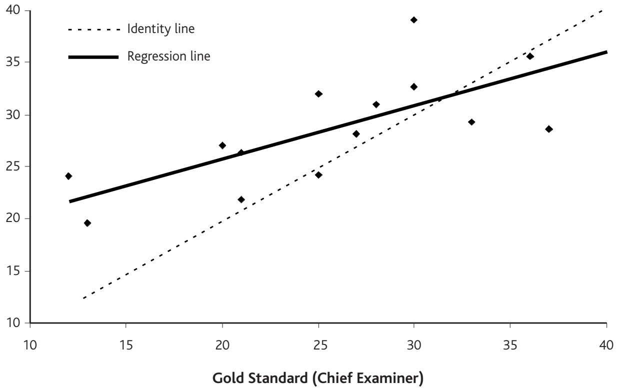


Figure 1: Comparison of CE's 'gold standard' total marks with mean teacher total marks

It can be seen that the regression line (bold line) is generally less tilted than the line of identity, showing that the teachers as a group tended to be less likely to use the extremes of the mark scheme than the CE. However, this could be interpreted as evidence of regression to the mean, as individually, the teachers used wider ranges of marks than the CE did.

Discussion

The above analyses indicate that marking reliability was good, though like almost all qualifications (Suto, Nádas and Bell, 2009), imperfect. Possible reasons and explanations for marking difficulty were identified by the participating teachers, which were recorded as written comments in their marking grids and questionnaire responses. Table 10 and Table 11 summarise the teachers' explanations for why some reports were harder and easier to mark than others.

The teachers' comments indicate that many of them found it easier to mark reports within their own subject areas, despite the generic nature of the Cambridge Pre-U IRR mark scheme. Subject knowledge appears to have facilitated some teachers' understanding of the language and terminology used. However, this experience was by no means universal, with one teacher commenting that clarity of thought was critical to marking ease, even with research reports on alien subject matter. Moreover, one teacher gave having 'too much subject knowledge' as a reason for finding some reports harder to mark than others. It may be that for this particular teacher, subject knowledge obscured his or her perception of generic skills. Other comments from the teachers point towards individual differences in perceptions of what affects marking difficulty: whilst one teacher felt that good performances were easier to mark, another teacher felt that poor performances were easier to mark.

Table 10: Perceived reasons for difficulty of marking some reports

Perceived reasons for finding some reports harder to mark than others	*Illustrative quotes from teachers*
Main reasons • Technical language and terms; lack of background/specialist knowledge (N = 8) • Density of language (N = 4) **Other reasons** • Evaluating quality of sources of information • Intellectually challenging • Discerning structure/arguments • Lack of proper evaluation • Too much subject knowledge	*"There was a lot of technical language upon which the arguments and analysis were based. One needed to keep all of these new technical terms in mind whilst trying to assess how effectively the sources and perspectives had been dealt with. It felt a bit like spinning plates, with constant shuffling from one part of the project to another to check for meanings and consistency of their use."* *"The critical thinking and evaluative aspects were tricky to pick out of the density of the text."* *"Not only was this far from my 'home area', but the terminology was foreign."*

Table 11: Perceived reasons for ease of marking some reports

Perceived reasons for finding some reports easier to mark than others	*Illustrative quotes from teachers*
Main reasons • Within subject area (N = 7) – taught – studied – familiarity – academic specialism • Clear analysis of perspectives; clarity of thought/argument/terminology (N = 5) **Other reasons** • Easy to judge use of source material • Short • Poor performance • Good performance • Marking familiarity – increased during course of study	*"...on a topic I have in-depth knowledge of."* *"It was easiest for me to mark the report on Communism as that is closest to my own academic specialism."* *"The ones which were easiest to mark were the reports presented with clarity of thought, even though the subject matter was unfamiliar to me."* *"...in my comfort zone of an area plus it was clearly argued and debated with discussion of the main criteria reflected in the mark scheme e.g. the notion of flaw etc."* *"...because it was easy to read, relatively short and at a low level."*

The teachers' comments provide a useful window into the nature of research report marking. However, it is worth noting that perceived marking difficulty is not the converse of marking accuracy. A marking task may feel difficult without accuracy necessarily being compromised, since assessors may put greater effort into demanding marking situations, as found by Johnson, Nádas and Bell (2009). Similarly, marking confidence may not be a good indicator of actual marking accuracy, since genuine insight into the marking process may be lacking, as has been found to be the case for some GCSE examiners (Nádas and Suto, 2007).

To conclude, the levels of marking reliability found in this study are encouraging. This is especially so given the study's limitations, which include the unavailability of authentic Cambridge Pre-U independent research reports, the novelty of the mark scheme, and the inexperience of the teachers involved in this study, who had no prior training and no access to material exemplifying standards. Future challenges for researchers include exploring assessment objectives that can only be assessed in the context of the classroom, by students' own teachers. Not all research skills can be assessed via a written research report and it is important that skills such as knowledge and understanding of the research process (AO1 in the Cambridge Pre-U's IRR mark scheme) can also be rewarded consistently.

Acknowledgments

We are very grateful to The Sixth Form College, Farnborough for the use of past students' project reports. We would also like to thank the teachers who participated in this study.

References

Brown, G., Bull, J. & Pendlebury, M. (1997). *Assessing student learning in higher education*. Routledge: London and New York.

Hatch, E. & Lazaraton, A. (1991). *The Research Manual: Design and Statistics for Applied Linguistics*. 533–535. Boston, Massachusetts: Heinle & Heinle.

Johnson, M., Nádas, R. & Bell, J. F. (2009). Marking essays on screen: An investigation into the reliability of marking extended subjective texts. *British Journal of Educational Technology*, published online.

Laming, D. (1990). The reliability of a certain university examination compared with the precision of absolute judgments. *Quarterly Journal of Experimental Psychology*, **42A**, 239–254.

Lunz, M.E. & Wright, B.D. (1997). Latent Trait Models for Performance Examinations. In: Jürgen Rost and Rolf Langeheine (Eds.) *Applications of Latent Trait and Latent Class Models in the Social Sciences*. http://www.ipn.uni-kiel.de/aktuell/buecher/rostbuch/ltlc.htm

McNamara, T.F. (1996). *Measuring Second Language Performance*. London: Longman.

Nádas, R. & Suto, I. (2007). An exploration of self-confidence and insight into marking accuracy among GCSE maths and physics markers. (A magabiztosság és a teljesítménybecslés pontosságának kutatása az angol GCSE vizsgák értékelo˝ inél') *Magyar Pedagogia*, **107**, 3, 169–184.

Suto, I., Nádas, R. & Bell, J.F. (2009). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education*, published online.

Shaw, S.D. (2008). Essay Marking On-Screen: implications for assessment validity. *E–Learning*, **5**, 3, 256–274.

Tomlinson, M. (2004). *14–19 Curriculum and Qualifications Reform: Final Report of the Working Group on 14–19 Reform*. Annesley, Nottinghamshire: DfES Publications.

Wright, B. & Linacre, J. (1994). Reasonable Mean-square Fit Values. *Rasch Measurement Transactions*, **8**, 3, 370.

Appendix: Stages in the two-day project report selection meeting

1. The CE read through all 346 project report titles and categorised them as either *no* or *yes/maybe* according to whether they would be verified as Cambridge Pre-U Independent Research Report titles. The main criterion for rejection was that a title did not explicitly invite discussion. Only titles that seem to lead to discussion are suitable for the IRR. There were 118 *yes/maybe* reports in total.
2. The CE and researchers looked at the participating teachers' subject backgrounds and made a list of all subject specialisms. Any missing major subjects (e.g. geography, psychology) were added to the list. The list was then revised and refined to form broad 16 subject areas, into which the project reports could probably be grouped.
3. The CE and researchers grouped the *yes/maybe* reports into the 16 subject areas.
4. The initial subject classifications of each report were checked, subject area by subject area, in a group discussion. Some reports were moved to different subject areas at this point. The numbers of reports in each subject area were counted (N = 118 in total).
5. The report titles in each subject area were checked (again in a group discussion) and the CE discarded any reports with titles that he did not think he could ultimately verify. The numbers of reports in each subject area were counted (N = 94 in total).
6. In a discussion of how to select the 20 reports needed for the study, the CE proposed that the criteria for spotting top reports would be:
   - Incisive conclusions (AO3)
   - Alternative interpretations (AO3)
   - Uses a range of sources (AO2)
   - Critical vocabulary (AO4)

   It was agreed that the first 5 reports that teachers mark should flag up key issues that need to be addressed in the CE's feedback.
7. The CE and researchers read through the reports in the subject areas (each taking the subject areas that s/he knew most about) and selecting one or two possible reports for inclusion in the sample on the basis of them being (1) very strong, (2) very weak, or (3) interesting and likely to generate discussion. This generated a selection of 23 reports.
8. Three reports from the most over-represented subject areas (economics, history and geography) were excluded from this selection to leave 20 reports.
9. The CE suggested that the practice sub-sample of reports should help the participating teachers to understand the marking criteria by illustrating key aspects of the mark scheme. The CE identified the following selection requirements:

   Report 1: AO2 and AO3 at level 3
   Report 2: AO2 and AO3 at level
   Report 3: AO2 and AO3 at level 1 or 2
   Report 4: AO5 at level
   Report 5: AO2, AO3, AO4 and AO5 at level 2.
10. In a group effort, five reports were found which met the above requirements and also covered a good mix of subject areas. They were then ordered so that they would not be encountered in either ascending or descending order of quality, but in a mixed order of quality.

Details of the selection process are summarised in Table A1.

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.

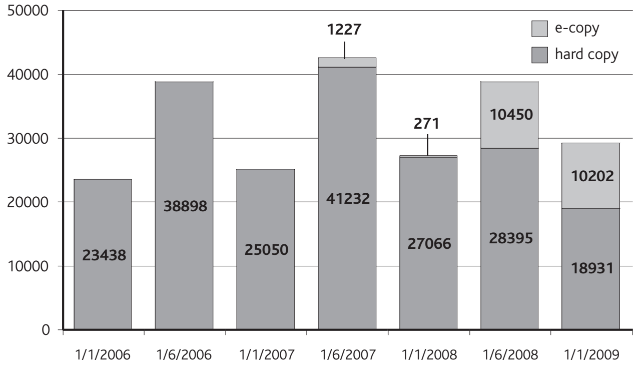
Context for the study

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).

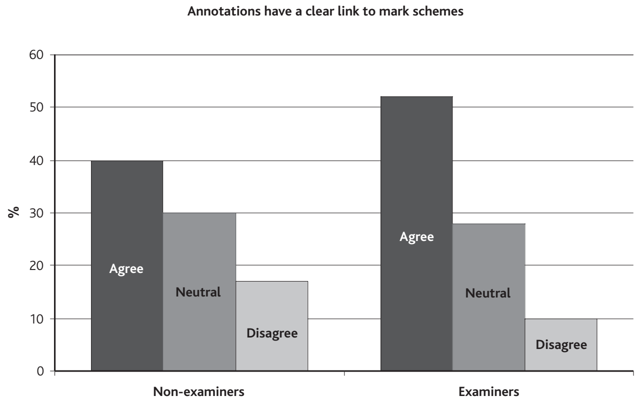


Figure 2: Perceived links between annotations and mark schemes

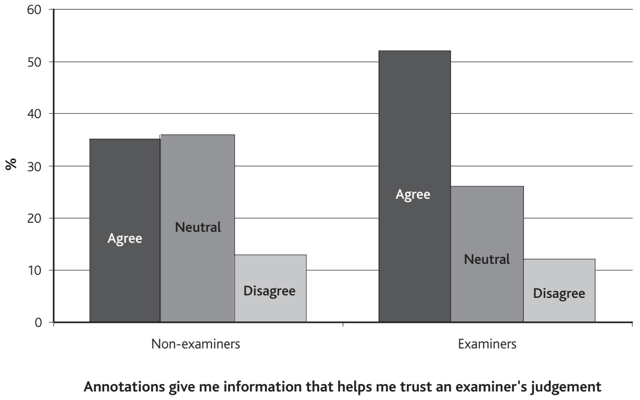


Figure 3: Annotations aid trust

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).

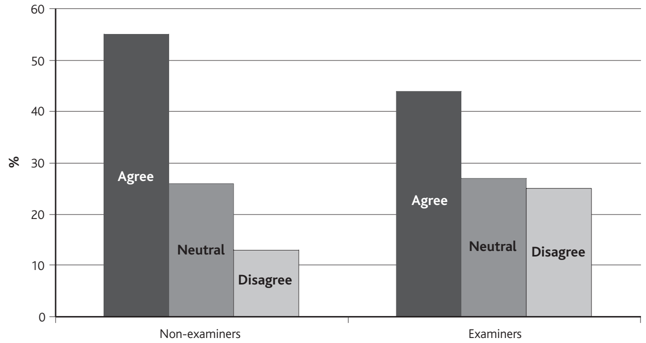


Figure 4: Annotations and formative purpose

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.

ASSURING QUALITY IN ASSESSMENT

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?

2. How does the effect on marking accuracy of a face-to-face meeting vary with the type of question being marked (short-answer or essay) and the level of experience of the examiners?
3. To what extent do examiners carry forward standardisation on one set of questions to a different but very similar set of questions?

We found that while standardisation improved marking accuracy for both new and experienced examiners, marking both short-answers and structured, factual essays, the benefit of including a face-to-face meeting in the standardisation process was variable, small and questionable. We also found that the effects of standardisation on one set of questions – with or without a meeting – carried forward into improved marking accuracy on other, very similar questions, implying that some transferable examiner learning had taken place and that the impact of – and need for – standardisation might decrease with examiner experience.

We concluded that it would be reasonable for examining bodies to explore whether standardisation can be achieved using more cost-effective and efficient methods than face-to-face meetings.

Background

The regulatory authorities for public examinations in England, Wales and Northern Ireland prescribe that awarding bodies must have a standardisation process that is "designed to make sure that all examiners mark candidates' work consistently and accurately [and which] establishes a common standard of marking that should be used to maintain the quality of marking during the marking period." (Qualifications and Curriculum Authority, 2009, section 4.14).

Research into the effectiveness of standardisation meetings has cast doubt on their usefulness, at least for experienced examiners. For example, Baird *et al*. (2004) found neither consensual meetings – where the examiners mutually agreed a common interpretation of the mark scheme – nor hierarchical meetings, where the Principal Examiner tried to impose his interpretation of the mark scheme on to the other examiners, improved the marking reliability of experienced GCSE History examiners. Similarly, Greatorex and Bell (2008) found that a standardisation meeting on its own had little effect on the reliability of experienced examiners of AS Biology. Greatorex *et al*. (2007) compared the pre- and post-standardisation meeting marking accuracy of experienced examiners of GCSE mathematics and physics with that of mathematics and physics graduates who lacked both teaching and examining experience and who would therefore not normally have been eligible to mark the examinations. They found that for the questions that the researchers had previously judged to entail more complex cognitive marking strategies, the standardisation meeting led to a much greater improvement of the graduates' accuracy than of the experienced examiners' accuracy. However, the improvement shown by graduates might also have occurred if other standardisation methods had been used, and might not be dependent on a standardisation *meeting* being held.

Method

Choice of examination

Two A-Level psychology units were chosen for the research, one assessed using short-answer questions, the other assessed using essay questions. We chose A-Level psychology because this subject uses both these types of question and because there is a large entry and correspondingly large pool of examiners.

Choice of examination questions

The short-answer examination we selected contained a number of discrete sections, each of which consisted of compulsory questions on a single topic. Two of the sections had identically structured questions, and by selecting these sections for the study and standardising examiners on only one of them, we could investigate the extent to which standardisation on one set of short answer questions carried over to other very similar questions answered by the same candidates. This would help us understand whether generic marking skills were developed through standardisation that lessened the impact of and need for standardisation in subsequent sessions, as examiners gained experience.

The essay examination gave candidates a choice of questions, so each question was answered by a different sub-group of candidates. We therefore investigated the carrying-forward of standardisation using essays from examinations held in consecutive years, selecting the closest matching questions for use in the study (question 4 in each case).

Some details concerning the chosen questions are given below:

Short answer questions

Questions which required candidates to write a sentence or two.

Short-Answer Collection 1 *Examiners were standardised on these*		*Short-Answer Collection 2* *Examiners were not standardised on these*	
Topic: Cognitive Psychology		Topic: Social Psychology	
Question	Mark tariff	Question	Mark tariff
1, 2a, 2b & 3 4	2 each 4	13, 14a, 14b, 15 16 4	2 each

Essay questions

Questions which required candidates to write a page or two of factual information.

Essay Collection 1 *Examiners were standardised on these*		*Essay Collection 2* *Examiners were not standardised on these*	
Examination 1		Examination 2	
Question	Mark tariff	Question	Mark tariff
4a, 4b	12 each	4a, 4b 12 each	

Participants

Twenty-four psychology examiners were recruited for the study, none of whom had operationally-marked the examinations. Twelve of the examiners had experience of marking other psychology A-Level examinations; the other twelve examiners were new to examining, having been recruited for operational work but not yet deployed.

The examiners were randomly assigned to experimental groups of six as follows:

	New Examiners	Experienced Examiners
Attends standardisation meeting	Group A1	Group B1
No meeting	Group A2	Group B2

In addition to these twenty-four examiners, two Team Leaders from the operational examinations were recruited, one from the short-answer examination, the other from the essay examination. These Team Leaders had each been responsible for supervising a team of examiners in the operational marking and were chosen based on the recommendations of the Principal Examiners and Professional Officer.

The role of the Team Leaders in the study was to standardise the other examiners and to provide reference marks for each answer against which the examiners' marks could be compared.

Overview of the sequence of events for Examiners

1. *Examiners marked pre-standardisation batches of scripts.*
   The marks from these scripts were used to calculate the examiners' pre-standardisation marking accuracies on each collection of questions (in relation to the Team Leaders' reference marks).
2. *Examiners were standardised, with or without a meeting according to their experimental group.*
3. *Examiners marked post-standardisation batches of scripts.*
   The marks from these were used to calculate the examiners' post-standardisation marking accuracies on each collection of questions (again in relation to the Team Leaders' reference marks).

Materials

Scripts

A random sample of scripts, stratified by grade, was drawn from the operational examinations once all marking and grading were complete.

The scripts were scanned and the marks and examiner annotations electronically deleted from the resulting images. The images relating to the questions chosen for use in the study were then printed out to give 'clean' hard copies. All participants marked the same answers, so twenty-six copies were printed.

The clean answers were divided into a number of batches, as shown below. The answers used in standardisation were selected by the Team Leaders. The pre- and post-standardisation batches were selected by the researchers and were matched by operational marks, so that the pre-and post- batches were as similar as possible.

Pre-standardisation batches:

Batch **Short-1-Pre** 50 answers to each question in Short-Answer Collection 1	Batch **Essay-1-Pre** 25 answers to each question in Essay Collection 1	Examiners were to be standardised on these questions
Batch **Short-2-Pre** 50 answers to each question in Short-Answer Collection 2	Batch **Essay-2-Pre** 25 answers to each question in Essay Collection 2	Examiners were **not** to be standardised on these questions

Batches for use in standardisation:
(Question collections 1 only. The standardisation procedure, described below, required three standardisation batches of each answer type)

Batch **Short-Stand-i** 5 answers to each question in Short-Answer Collection 1	Batch **Essay-Stand-i** 5 answers to each question in Essay Collection 1
Batch **Short-Stand-ii** 5 answers to each question in Short-Answer Collection 1	Batch **Essay-Stand-ii** 5 answers to each question in Essay Collection 1
Batch **Short-Stand-iii** 10 answers to each question in Short-Answer Collection 1	Batch **Essay-Stand-iii** 10 answers to each question in Essay Collection 1

Post-standardisation batches:

Batch **Short-1-Post** 50 answers to each question in Short-Answer Collection 1	Batch **Essay-1-Post** 25 answers to each question in Essay Collection 1	Examiners were standardised on these questions
Batch **Short-2-Post** 50 answers to each question in Short-Answer Collection 2	Batch **Essay-2-Post** 25 answers to each question in Essay Collection 2	Examiners were **not** standardised on these questions

Materials written by the Team Leaders

The Team Leaders were commissioned to write:

- an *Introduction to Marking* for new examiners;
- a *Mark scheme Rationale* explaining to examiners how the mark schemes for the chosen questions should be applied;
- written explanations for the marks they awarded to the first and second standardisation batches of short answers and essays. Copies of these would be placed in sealed envelopes for the examiners to open and read when directed, as described below under 'Experimental Procedure'.

Additional materials supplied to participants

- Copies of the question papers
- Copies of the relevant parts of the mark schemes
- Instructions

Experimental Procedure

Stage 1: Pre-standardisation

(1) The pre-standardisation batches were posted to the examiners, together with copies of the questions and mark schemes.

(2) Examiners were instructed to mark the pre-standardisation batches in the following order: Short-1-Pre first, then Essay-1-Pre, then Short-2-Pre, then Essay-2-Pre.

(3) Examiners returned their marked pre-standardisation batches.

(4) The remaining materials were posted to examiners.

Stage 2: Standardisation
The standardisation procedure was the same for all examiners, except for the inclusion of a standardisation meeting for examiners in experimental groups A1 and B1.

	Groups A1 & B1	*Groups A2 &B2*
(5)	All examiners were instructed to read *Introduction to Marking* and the questions, mark schemes and mark scheme rationale.	
(6)	All examiners marked batch Short-Stand-i, then opened the envelope containing the Team Leader's marks and explanations for Short-Stand-i. They were instructed to compare the Team Leader's marks with their own and read the explanations.	
(7)	All examiners marked batch Short-Stand-ii.	
(8)		A2 & B2 examiners opened the envelope containing the Team Leader's marks and explanations for batch Short-Stand-ii. They were instructed to compare the marks with their own and read the explanations.

(9)	All examiners marked batch Essay-Stand-i, opened the envelope containing the Team Leader's marks and explanations, compared the marks with their own and read the explanations.	
(10)	All examiners marked batch Essay-Stand-ii.	
(11)		A2 & B2 examiners opened the envelope containing the Team Leader's marks and explanations for batch Essay-Stand-ii. They were instructed to compare the marks with their own and read the explanations.
(12)	A1 & B1 examiners attended a standardisation meeting, at which their marking of Short-Stand-ii and Essay-Stand-ii was discussed and the correct marks provided and explained. At the end of the meeting the examiners were also supplied with copies of the written explanations and marks previously given to the non-meeting groups, so that all had the same materials.	
(13)	All examiners marked batches Short-Stand-iii and Essay-Stand-iii. They were instructed to enter their marks into spreadsheets and email them to the appropriate Team Leader.	
(14)	Team Leaders phoned each examiner individually to discuss their Stand-iii marking and answer questions.	

Stage 3: Post-standardisation

(15)	Examiners marked the post-standardisation scripts in the following order: Short-1-Post first, then Essay-1-Post, then Short-2-Post and finally Essay-2-Post.
(16)	Examiners returned all their marked scripts.

Additionally, the Team Leaders marked the pre- and post-standardisation batches to provide reference marks for use in the analysis. Each Team Leader marked only short answers or essays according to their specialism.

The standardisation meeting

Examiners in groups A1 and B1 attended a standardisation meeting in Cambridge, led by the two Team Leaders. After a preliminary welcome, a brief presentation was given by one of the Team Leaders recapping the material contained in the *Introduction to Marking* document. Consecutive sessions were then held for the short-answer and essay questions, each led by the appropriate Team Leader and conducted as similarly as possible to the operational standardisation meeting. During these sessions examiners went through the second standardisation batches and the Team Leader led a discussion of the examiners' initial marks and provided and explained the 'correct' marks. Examiners had ample opportunity to ask questions.

Analysis

The 'absolute difference' between each examiner's mark for an answer and the reference mark was calculated – this was simply the value obtained by subtracting examiner-mark from reference-mark and discarding the sign, that is, all were positive numbers. These absolute differences gave the size of the difference, and when averaged did not cancel out as actual differences might.

The mean absolute difference was calculated for each examiner on each question in the pre- and post-Standardisation collections. Means were also calculated at the level of experimental group, and batch.

Analysis of covariance (ANCOVA) was performed to test whether post-standardisation differences between the experimental groups were statistically significant, having controlled for pre-standardisation differences.

Results and discussion

The charts in this section show the pre- and post-standardisation mean absolute-difference between examiner-mark and reference-mark for each experimental group. The solid lines correspond to the results from the examiners who attended the meeting ('Face-to-face' standardisation type), the dotted lines to those from the examiners who did not attend the meeting ('Remote' standardisation type). Statistical significance information from the ANCOVA analyses are given underneath the charts, where ✓ indicates $p < 0.05$, i.e. where examiner experience, or standardisation type, or different combinations of these two factors ('interaction') resulted in statistically significant differences in post-standardisation absolute-differences.

The first thing to note from the charts is that in almost all cases standardisation had a beneficial effect in bringing examiners' marks closer to the reference marks, regardless of whether examiners attended the meeting. The ANCOVA analysis helps determine whether meeting attendance had an *additional* effect on marking accuracy, over and above that derived from undertaking the remote standardisation tasks, and whether this varied with examiner experience.

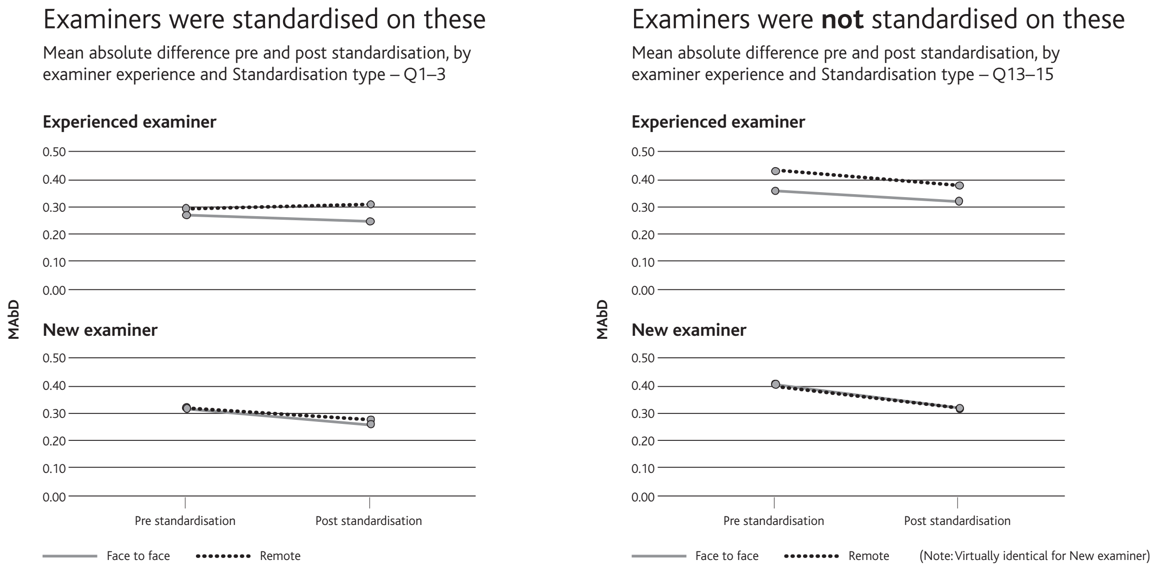
Short-answer questions

Figure 1 shows the pre- and post-standardisation mean absolute-differences for each experimental group on the 2-mark questions. The charts on the left show the results on the standardised questions, those on the right give the results on the un-standardised questions. In both cases the experienced examiners' results are presented in the top charts.

There was a slight but statistically significant benefit (in terms of reducing mean absolute differences) in attending the standardisation meeting for the standardised questions only. For the un-standardised questions, attending the meeting did not provide a general significant benefit, but there was a significant but very small interaction between standardisation type and examiner experience: from the diagrams it is apparent that there is no difference between the lines for the new examiners, but those for the experienced examiners are a little less than parallel.

Figure 2 shows the results for the 4-mark question. Clearly standardisation had unintended consequences for question 4: marking accuracy worsened! This is the only question for which this is the case. Examiner experience had a significant effect, with the experienced examiners' accuracy worsening slightly less; attending the meeting had a particularly negative effect on the new examiners. On question 16, the 4-mark question on which examiners were not standardised, meeting attendance resulted in a very slight, but statistically significant, improvement.

Figure 1: 2-mark questions


Examiners were standardised on these

Mean absolute difference pre and post standardisation, by examiner experience and Standardisation type – Q1–3

Examiners were **not** standardised on these

Mean absolute difference pre and post standardisation, by examiner experience and Standardisation type – Q13–15

(Note: Virtually identical for New examiner)

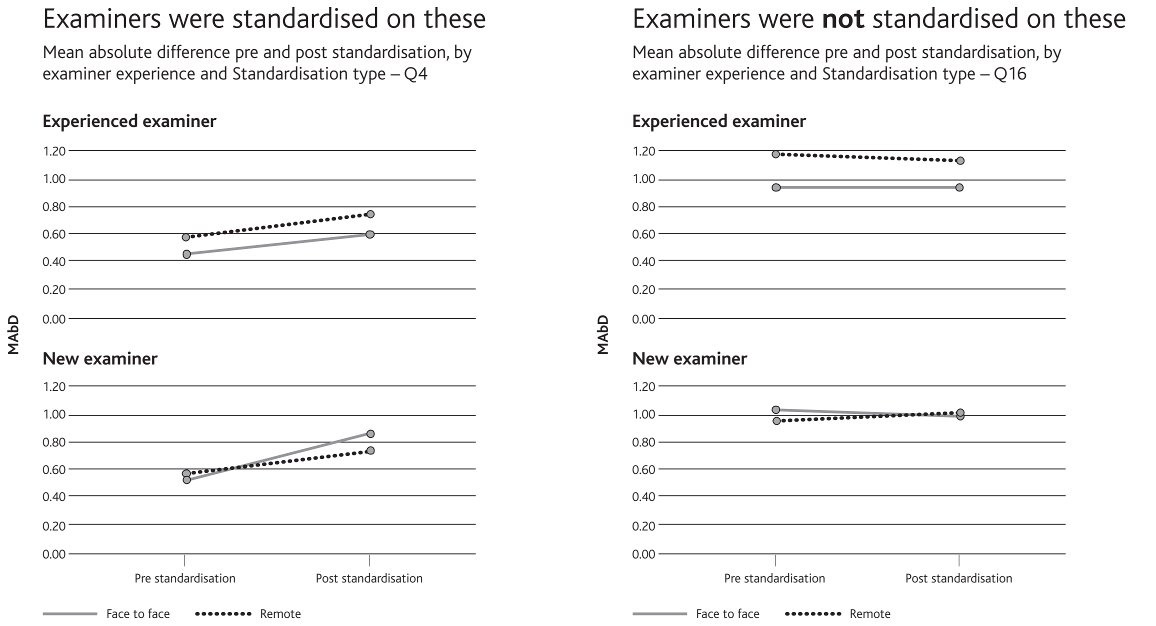
Statistical significance

Examiner experience	✗	p=.710
Standardisation type	✓	p=.003
Interaction	✗	p=.138

Statistical significance

Examiner experience	✗	p=.096
Standardisation type	✗	p=.084
Interaction	✓	p=.044

Figure 2: 4-mark question


Examiners were standardised on these

Mean absolute difference pre and post standardisation, by examiner experience and Standardisation type – Q4

Examiners were **not** standardised on these

Mean absolute difference pre and post standardisation, by examiner experience and Standardisation type – Q16

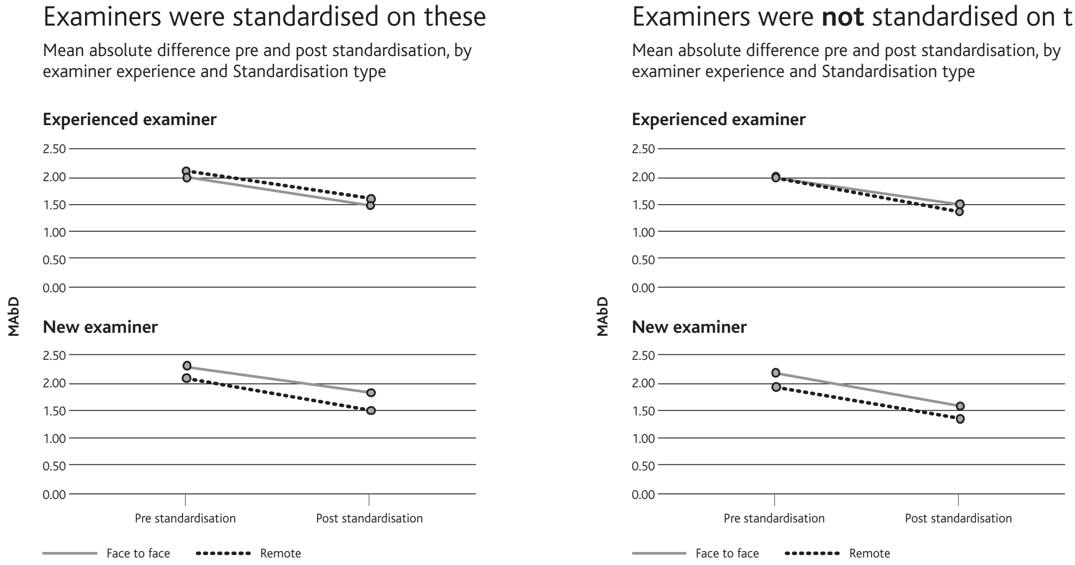
Statistical significance

Examiner experience	✓	p=.002
Standardisation type	✗	p=.947
Interaction	✓	p=.002

Statistical significance

Examiner experience	✗	p=.934
Standardisation type	✓	p=.040
Interaction	✗	p=.135

Figure 3: Essay questions



Statistical significance

Examiner experience	✘	p=.094
Standardisation type	✘	p=.282
Interaction	✔	p=.008

Statistical significance

Examiner experience	✘	p=.745
Standardisation type	✔	p=.045
Interaction	✘	p=.795

Figure 3 gives the results for the essay questions. Standardisation was clearly beneficial on both the standardised and non-standardised questions. Neither standardisation type nor examiner experience had a significant effect on the accuracy improvement on the standardised questions, but there was a significant interaction between these factors, with the remotely standardised new examiners improving more. On the un-standardised questions there was a statistically significant greater improvement for the remotely standardised examiners, with the chart suggesting that this greater improvement was shown mainly by the experienced examiners, though the interaction between experience and standardisation type was not significant.

Conclusions

On the basis of our results, we concluded that:

- Apart from the anomalous 4-mark question, standardisation improved the examiners' marking accuracy when compared with the reference marks, regardless of whether this standardisation was conducted purely remotely or with the addition of a face-to-face meeting.
- The standardisation improvement carried over into other, very similar questions, implying the examiners learnt lessons from being standardised that they were able to apply when marking other questions. This finding suggests the impact of – and need for – standardisation might reduce with examiner experience.
- Meeting attendance did not always have a statistically significant benefit, and where there was a benefit, it was very small in real terms. On the standardised questions, the meeting yielded a significant benefit on the 2-mark questions, but not on the essays, where the remotely standardised new examiners improved more than those attending the meeting. On the un-standardised essay questions, remotely-standardised examiners improved more than the meeting attendees.

From the perspective of improving marking accuracy in relation to Team Leader reference marks, the benefits of holding a face-to-face standardisation meeting therefore appear variable, small and questionable, for both new and experienced examiners, and for both essay and short-answer questions. It would be reasonable for examining bodies to explore whether standardisation can be achieved using more cost-effective and efficient methods than face-to-face meetings.

Caveats

A number of caveats must be placed on these findings.

- The essays were highly structured and factual, and marked against a prescriptive mark scheme. Findings might not be replicated with less constrained essays and marking.
- The Team Leaders were not experienced at leading standardisation, a task carried out operationally by the Principal Examiner. They were recommended to us for this task, however.
- We used only two Team Leaders, one for short-answers, the other for essays. We therefore have no way of separating any effects introduced by the Team Leaders from effects introduced by the question type. Similarly, each reference mark was produced by only one Team Leader, who may or may not have been typical – though the fact that both had been successful Team Leaders in the operational marking mitigates against this risk.
- Only twenty-four examiners took part in the study, and these examiners might not have been representative of the wider populations of experienced and new examiners.
- Both the meeting and the remote standardisation tasks differed from normal operational practice. Cambridge Assessment only uses remote standardisation methods in the context of online marking,

where examiners can be monitored and supported more effectively than when marking on paper. In the present study all marking was carried out on paper, and the standardisation tasks adapted to match as closely as possible with those used operationally with online marking. Operational standardisation meetings are conducted by Principal Examiners and focus on either the short-answer examination or the essay examination, but not both. Examiners typically mark only one examination. However, the number of questions used in the study was far fewer than would be used in an operational setting.

- All participants knew that the marks did not 'count', and were only for use in the research. Whilst it is our impression that all participants were highly diligent and professional, we have no way of quantifying what effects, if any, were introduced by the low stakes nature of the exercise.

Finally, it should be noted that in operational marking settings examiners are given additional standardisation if necessary and are removed from the marking panel if their accuracy remains unsatisfactory. Additionally, examiners' operational marking is sampled on several occasions after initial standardisation, to check that accuracy levels are maintained. For these reasons operational marking is likely to be more accurate than was found in this study.

References

Baird, J., Greatorex, J. & Bell, J.F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education*, **11**, 3, 331–348.

Greatorex, J. & Bell, J.F. (2008). What makes AS marking reliable? An experiment with some stages from the standardisation process. *Research Papers in Education*, **23**, 3, 333–355.

Greatorex, J., Nádas, R., Suto, I. & Bell, J.F. (2007). *Exploring how the cognitive strategies used to mark examination questions relate to the efficacy of examiner training*. Paper presented at the ECER conference, Ghent, Belgium in September 2007.

Qualifications and Curriculum Authority (March 2009). *GCSE, GCE and AEA Code of Practice*. London: QCA.

ASSURING QUALITY IN ASSESSMENT

A review of literature on item-level marker agreement: implications for on-screen marking monitoring research and practice

Milja Curcin Research Division

Introduction

Marking reliability contributes in important ways to the overall reliability and validity of assessment. It refers to the extent to which different examiners' marks agree with each other or with a definitive mark when they mark the same material (inter-marker agreement), and is also affected, for instance, by individual examiners' consistency throughout marking (intra-marker consistency). Validity of assessment is compromised without high marking reliability since the same mark from different examiners cannot be assumed to mean the same thing (e.g. Massey and Raikes, 2006; Cambridge Approach, 2009). However, as Wilmut *et al*. (1996) observe, "[f]or a variety of reasons, perfect reliability is not going to happen. The aim must be to get as close as possible, given irreducible constraints."

This review article focuses mainly on the literature relevant for the inter-marker agreement aspect of marking reliability in the context of on-screen marking. The increasing use of on-screen in place of paper-based marking presents new possibilities for monitoring of marking and ensuring higher agreement levels, but also raises questions with respect to the most efficient and beneficial use of marker agreement information that is routinely collected in this process, both in monitoring practice and in research.

Current Ofqual[1] regulations (Code of practice, April 2009) for on-screen marking require that the marking of individual examiners be compared to that of a senior examiner at regular intervals throughout the marking process. Although the specifics of this procedure differ across awarding bodies, this is generally implemented by means of "seeding" pre-marked "seeding scripts" (or items)[2] into live marking at regular intervals. The markers' marks are checked against the scripts'/items' "definitive marks",[3] these having been determined in advance by a single senior examiner or by a panel of senior examiners, depending on awarding body practices.

In this monitoring process, marker agreement data are collected at item level, potentially providing a rich source of information, particularly with respect to which features of items are associated with high or low marker agreement. Furthermore, since some awarding bodies use expert panels to decide on definitive marks, presumably under the assumption that groups make better decisions than individuals (cf. Levine and Moreland, 2006), it is conceivable that the group dynamics of these panels could affect the choice of the definitive marks and subsequent individual marker agreement with them. It is useful, therefore, to consider research to date on marker agreement, particularly at item level, as well as social psychology research on group dynamics, as this might inform both current marking monitoring processes and future research in this area, particularly in respect of what marker agreement levels can be

[1] Ofqual (Office of the Qualifications and Examinations Regulator) is responsible for regulating public examinations.

[2] Script: whole candidate work on one question paper. Item: candidate response on one question or question part.

[3] The definitive marks are not visible on the scripts.

expected in different assessment contexts and with different assessment types.

The article first briefly reviews several studies into marker agreement at script level, focusing subsequently on research investigating finer-grained factors affecting agreement at item level, particularly with respect to marking task demands. This is followed by a brief overview of research into group dynamics and small-group decision making relevant to the group dynamics in expert panels deciding on definitive marks.

Marker agreement at script level

Most marking reliability studies conducted before the rise of on-screen marking have been conducted at whole script level, partially replicating common marking monitoring practices in paper-based marking. In several experimental studies in this context, Murphy (1978, 1979, 1982) used blind[4] re-marking to investigate mark/re-mark agreement. Overall, for nearly all of the 20 different GCE O and A-level examinations that were investigated, the correlation coefficients comparing prime and re-mark were above 0.90, except for English, where they were between 0.73 and 0.93 for individual papers (between 0.80 and 0.95 for combined papers). More recently, Massey and Raikes (2006) conducted a blind multiple-marking study on sample items taken from GCE A-Level and IGCSE examinations in a range of subjects and reported on intraclass correlations (ICCs)[5] at paper level for each subject. The ICCs they reported were in the range of 0.77 (Economics) to 0.99 (French).

The usual monitoring procedures in paper-based marking, however, involve a senior examiner re-marking a sample of each of their team's allocation of scripts at several points in the marking process, while re-marking is non-blind. Pinot de Moira *et al*. (2002, on A-level English) and Bramley (2008, on 38 different subjects) investigated mark/re-mark agreement data collected as part of such monitoring process. They found that mark/re-mark correlations generally exceeded 0.95. However, both studies acknowledge that non-blind re-marking may have boosted marker agreement. Indeed, Murphy (1979) and a number of other studies (e.g. Wilmut, 1984; Massey and Foulkes, 1994; Vidal Rodeiro, 2007) have demonstrated that inter-examiner agreement tends to be lower when the re-marking process is blind.

Importantly, most studies reviewed above report somewhat different agreement levels for different subjects. Murphy's (1978) findings also indicated that question type is an important factor, as suggested by different levels of agreement on differently structured papers within, for example, Geography O-level and English A-level, where papers with more structured questions had higher mark/re-mark correlations. This is further demonstrated in his 1982 study, where he noted that the examining technique (i.e. using essay-type vs. objective questions) tended to outweigh between-subject differences. These findings were replicated by Newton (1996) for English and Mathematics.

Clearly, investigating marker agreement at script level rather than at item level makes it difficult to separate the relative effect on marker agreement of various fine-grained factors including question type. The following section reviews studies that investigate marker agreement at item level mainly in the context of on-screen marking, which attempt to establish relative importance of these different factors and determine the operational potential and value of controlling for at least some of them in order to increase marker agreement in problematic areas.

Fine-grained features affecting marker agreement

Factors affecting marker agreement can be grouped into two general categories, depending on whether they reside in the demands of the marking task or in the marker's personal expertise (see Black, Suto and Bramley, *in submission*). The first group of factors includes item features, mark scheme features, and candidate response features. Some of the prominent factors residing in the marker include expertise, level of education and amount of training. This review will focus on the first group of factors as they are particularly relevant in the context of on-screen marking monitoring by means of seeding items in that they might inform the choice of seeding items and predictions regarding where marker agreement might be low or high.

Since in some awarding bodies (e.g. OCR[6]), the definitive marks of the seeding items are agreed by an expert panel of senior examiners, the group dynamics of these panels could be expected to interact in complex ways with factors related to the marking task and affect the choice of the definitive marks as well as subsequent marker agreement with these marks. A separate section below is therefore dedicated to an overview of research dealing with small group decision making and group dynamics.

Item and mark scheme features

One of the first studies specifically designed to investigate how different features of marking task could affect marker agreement at item level was Massey and Raikes (2006, see previous section), who investigated several surface features of items and their mark schemes (subject; maximum mark available for item; implied time restriction for candidates; type of marking: objective, points-based or levels-based; and number of levels available for levels-based marking).

Their results were mixed. Overall mean ICCs were the highest for objective items (0.97), next highest for points-based items (0.82) and lowest for levels-based items (0.77). On average, agreement decreased with rising maximum mark for points-based items, but this trend was unexpectedly reversed for Chemistry. Another interesting finding was that Sociology essay questions marked against a levels-based mark scheme were marked very reliably (average ICC=0.83, with little variation between items), indicating that it is possible to mark longer pieces of work using less constrained mark-schemes quite reliably. In general, although indicative of interesting patterns in terms of item type and other effects, these findings called for further study on larger quantities of data, and, as suggested by Suto and Nádas (2008), potentially indicate the need for a more sophisticated system of classifying questions according to marking demands.

Hudson *et al*. (2007) investigated on-screen marking reliability on seeding items for nine papers from three AQA[7] subjects. They investigated various factors, including: item type; item maximum mark; number of times the examiner had previously seen the same seed; at what time of day the marking was done. The effects of the first two factors are particularly relevant for inter-examiner agreement, and thus

[4] In blind re-marking, the examiners who re-mark cannot see the original markers' marks.

[5] Statistic describing how strongly units (in this case, marks) in the same group resemble each other, thus an indicator of examiner agreement.

[6] OCR (Oxford, Cambridge and RSA) is one of three main awarding bodies in England.

[7] AQA (Assessment and Qualifications Alliance) is one of three main awarding bodies in England.

for this review. Item type was defined in terms of whether an item could be marked by a (i) 'general' marker who was not a subject expert, or (ii) by a subject expert. Clearly, this definition conflates several item properties that could potentially be dissociated (e.g. expected response type, mark scheme properties, etc.).

Regarding item maximum mark, their findings replicate the findings elsewhere in the literature that higher tariff items tend to have higher absolute mark differences between definitive and examiner mark. However, the findings regarding item type (as defined in this study) are less clear-cut. In some subjects, the expert items tended to be associated with lower absolute mark differences, while in others this was the reverse. The authors acknowledge that there is probably a complex relationship between item type, marker expertise, marking variability and seed tolerances. Similarly to the Massey and Raikes (2006) study, it is clearly necessary to identify finer-grained distinctions when classifying item types for the purpose of marking reliability investigation.

Bramley (2008) attempted to identify some of these finer distinctions and coded a number of salient features of items and their mark schemes in order to investigate the relationship of the coded features with the level of marker agreement. The study made use of a large database of marker agreement data collected as part of the usual non-blind re-marking process at item (i.e. sub-question) level in June 2006 (OCR) and November 2006 (CIE[8]) from 38 subjects. The features coded included item maximum mark; item type (here defined in terms of whether the mark scheme was objective, points-based or levels-based); the amount of space available to the candidate to present their answer; the amount of writing required; the ratio of acceptable answers (points) allowed by the mark scheme to the number of marks available (points/marks ratio); whether the mark scheme specified qualifications, restrictions or allowable variants to the creditworthy responses; and whether the mark scheme specifically identified wrong answers.

The study used exact agreement (P_0)[9] as the measure of marker agreement, and logistic regression modelling to estimate the size and significance of the effect of coded features on this statistic. All the features were shown to be associated with marker agreement to a greater or lesser extent. However, three features were found to account for most of the explainable variance in marker agreement on objective and points-based items worth up to 9 marks. These were the number of marks available for the item, item type (objective vs. points-based), and the points/marks ratio. These features affected marker agreement in the expected direction: lower tariff, more constrained items with the number of acceptable answers equal to the number of marks had the highest agreement. In general, as Bramley observes, these findings fit the expectation that the amount of constraint in the mark scheme affects the marking accuracy and agrees with the findings of Massey and Raikes (2006) and other studies reviewed in this section.

A comparison of the relative influences of points-based vs. levels-based items did not yield clear-cut results though, that is, exact agreement was actually higher for levels-based items above 10 marks, perhaps contrary to expectation. Although this finding needs further investigation, Bramley suggests that a more 'subjective' mark scheme will not always necessarily lead to less accurate marking (cf. Massey and Raikes, 2006). Another possible explanation is that the re-marking in this study was non-blind, which may have affected the reliability patterns observed (cf. Black, Curcin and Dhawan, *in submission*, below) and also might have caused higher overall levels of agreement than would be expected in a blind re-marking situation (see previous section).

Influence of some of the above-mentioned features was also detected in the studies by Suto and Nádas (2008; 2009) investigating how examiners' thinking and their marking accuracy are affected by marking task demands defined in terms of cognitive marking strategy complexity (Greatorex and Suto, 2006; Suto and Greatorex, 2008a, b). Suto and Nádas (2008) found a strong relationship between the *apparent* cognitive marking strategy complexity (coded by researchers)[10] and marker agreement. While such findings obviously have practical implications in terms of allocating "simple-strategy" questions to general markers, and "complex-strategy" questions to expert markers, Suto and Nádas (2009) point out that it may not always be straightforward to categorise questions in terms of a relatively abstract characteristic such as marking strategy complexity.

In Suto and Nádas (2009), expert examiners used Kelly's Repertory Grid technique to identify the most influential features of questions that in their view contribute to marking strategy complexity. They identified about ten relevant features, five of which were particularly likely to demand the use of complex marking strategies and affect marker agreement: complexity of the candidate's presentation of ideas; amount of careful reading; independent vs. follow-through marks; use of words/formulae by candidate; whether the question involves application or recall of ideas; and scope/range of acceptable answers (i.e. points/marks ratio, cf. Bramley, 2008). All these features were identified as relevant for at least one subject (Biology, Mathematics or Physics) by Suto and Nádas (2008). In addition, Suto, Nádas and Bell (2009) found that the most important predictors of marker agreement for more complex strategy items were: target grade (reflecting predicted difficulty of question for candidate) and total mark (i.e. maximum mark, see for example, Bramley, 2008; Massey and Raikes, 2006).

In another study specifically designed to investigate marker agreement on seeding items[11] (Black, Curcin and Dhawan, *in submission*; see also Black, Suto and Bramley, *in submission*), data were collected on the seeding items used in the January 2009 session for five OCR units marked online in scoris®.[12] This study combined the insights from several studies cited above in terms of a comprehensive list of item/mark scheme features investigated. Most importantly, item type was defined more precisely in terms of level of constraint (objective, constrained, short answer question, extended response) while the mark scheme approach was defined separately as either objective, points-based or levels-based. Other features coded included maximum mark, definition of outcome space (whether the mark scheme specifies an exhaustive list of creditworthy responses or not), apparent marking strategy complexity (AMSC), physical answer space, whether wrong answer was specified, etc.

The features which were most strongly associated with differing levels of exact marker agreement were item maximum mark (the higher the tariff, the lower the agreement), item type (the more constrained the item, the higher the agreement), mark scheme approach (again, more constraint leads to higher agreement), definition of outcome space (the more exhaustive the outcome space, the higher the agreement), and AMSC (simple strategy – higher agreement). Thus, this study replicated

[8] CIE (University of Cambridge International Examinations) – another awarding body, providing international qualifications.

[9] The proportion of cases with no difference between a marker's mark and the definitive mark.

[10] The categorisation of marking strategy complexity in this study was based on researcher rather than examiner judgement, hence the *apparent* marking strategy complexity.

[11] Using the P_0 statistic (cf. Bramley, 2008).

[12] Bespoke software for online marking, developed by RM on behalf of OCR.

some of the important findings from previous research in this domain while providing further evidence for the influence of some previously un(der)explored factors.

Black, Suto and Bramley (*in submission*) suggest that question type and mark scheme approach may be key determining factors of cognitive marking strategy complexity, which they characterise as a fundamental concept that embodies various factors affecting the demands of the marking task and consequently marker agreement. Though question type and mark scheme approach seem indeed to be relevant, there are also other factors that can potentially make an apparently simple strategy question complex to mark for any particular marker. In particular, as noted in Bramley (2008), the difficulty with applying cognitive marking strategy complexity categorisation in advance in order to predict marker agreement (e.g. by researchers, or awarding bodies) is that the actual strategy applied in each case will depend to some extent on what the candidate has actually written. Irrespective of how much constraint is placed on the outcome space, candidates can always respond in an unanticipated fashion thus potentially affecting marking task demands and subsequent marker agreement.

Candidate response features

A number of studies have investigated the features of candidate responses that potentially influence examiners' choice of marks, both in marking and grading contexts. The majority of these features appear to be 'relevant' for the construct that is assessed in any particular subject, but there are also those that may not be, but still might affect examiners' judgement and marks (e.g. Crisp, 2007).

For instance, several experimental studies detected an influence of handwriting neatness and legibility on the marks awarded, with neater responses getting higher marks. The majority of these studies were conducted in experimental settings, where teachers were marking scripts of the same content but written in different handwriting styles (e.g. Briggs, 1970, 1980; Bull and Stevens, 1979; Markham, 1976). Massey (1983), however, failed to detect a significant influence of several potentially construct-irrelevant response features on marks given in A-level English literature exams in a study using a sample of actual marked scripts. He investigated the effect of features such as untidiness, prose complexity and prose accuracy on marks awarded. He suggests that the reason why this study failed to replicate previous findings might be that the markers in earlier studies were teachers, while the markers in this study were experienced examiners. The latter, through their procedures and/or experience, might be less likely than teachers to be influenced by candidates' writing. Another possibility is that there are differences in how markers of different subjects deal with different penmanship styles, or that handwriting and style differences are less pronounced the older the candidates are (i.e. A-level vs. GCSE).

Black, Curcin and Dhawan (*in submission*) also investigated the effect of some candidate response features on marker agreement, namely spelling, communication, legibility of handwriting, crossings-out, whether the response was standard or not, and whether it was in designated response area. Spelling, legibility and quality of communication were found to have only small effect on marking agreement, corroborating to some extent the findings of Massey (op. cit.).

Response features found to be most strongly associated with P_0 in this study were whether the response was standard (associated with higher agreement); the presence of crossings out (associated with lower agreement); and whether the response was entirely in the designated response area (associated with higher agreement). The latter two effects were characterised as unexpected since they are relatively superficial aspects of responses that should not increase the demands of the marking task. If indeed replicable, the latter effect in particular should probably be taken seriously considering the preponderance of out-of-area responses in candidate scripts (cf. Whetton and Newton, 2002). Furthermore, Black, Suto and Bramley *(in submission)* report that these last three features interact with other features of the marking task, in particular question type, mark scheme approach and AMSC, increasing the demand of the marking task even for some apparently simple marking strategy questions.

Group dynamics in expert panel decisions about definitive marks

According to Suto and Greatorex (2008a, b), from a cognitive psychological perspective, the individual judgements made in examination marking may not be fundamentally different from those made in other decision-making situations. However, since the decisions about definitive marks for seeding items are sometimes made by expert panels rather than individual examiners, usually by small groups of examiners led by one most senior examiner, these decisions can be seen as additionally subject to the influence of various social factors, for example, group polarisation (Fitzpatrick, 1989), minority influence (Brennan and Lockwood, 1980), the influence of 'authority' figures or personalities, and social conformity (Murphy *et al*., 1995).

Conformity, cohesion and dissenting minorities

A number of studies have investigated the impact of majority influence or conformity in group decision-making, observing that in many cases individuals change their opinions when they find out what is the majority opinion in their group (e.g. Asch, 1951, 1956; Deutsch and Gerard, 1955), and that this can be problematic if the majority opinion is misguided. Conformity in turn can lead to group polarisation. This refers to an initially dominant position becoming more extreme or enhanced as a result of group discussion (Moscovici and Zavalloni, 1969; Myers, 1982, cited in van Avermaet, 1988) which can sometimes lead to group-think, an extreme example of group polarisation (Janis, 1972, cited in van Avermaet, 1988).

According to Kerr and Tindale's review (2004), several recent meta-analyses indicate that more cohesive groups tend generally to be more productive if their group norms favour high productivity and their group members are committed to performance goals. However, high cohesion (and/or conformity) can also cause the loss of the beneficial effects of dissenting minorities (Zimbardo and Leippe, 1991, cited in Murphy *et al*., 1995). Several studies have shown that the presence of a dissenting minority can improve the quality of group decisions through greater consideration of alternatives, divergent thinking, and integration of multiple perspectives (e.g. Moscovici, 1976). This however, depends on a number of factors, particularly in situations when there is no demonstrable correct solution to a problem under discussion, for instance, to what extent the minority members are actually aware of the superiority of their opinion or knowledge (Phillips and Lewin Loyd, 2006).

Leadership styles and group performance

In some decision-making situations, groups may be organised in such a way that multiple people provide advice to a decision maker, but the final decision is in the hands of a single person. This corresponds to the set-up of expert panels deciding on definitive marks. Kerr and Tindale (2004) discuss a line of research dealing with these “judge-advisor systems” (e.g. Budescu and Rantilla, 2000; Sniezek, 1992, cited in Kerr and Tindale, 2004) and review a number of studies investigating how much influence the “advisors” have on the final decision of the “judges” (e.g. Harvey *et al*., 2000; Budescu *et al*., 2003). A general finding is that advisors influence judges, but judges give their own positions more weight and they also give more weight to advisors whose preferences are similar to their own, or who have been right in the past. However, the best predictor of an advisor’s influence appears to be his/her (apparent) level of confidence.

Another line of research deals with leadership styles, distinguishing democratic from autocratic leadership (e.g. Lewin and Lippitt, 1938; Lewin *et al*., 1939, cited in Gastil, 1997). As summarised by Gastil (1997), in the former case the leaders encourage group decision-making and discussion, active member involvement, honest praise and criticism, and a degree of comradeship. By contrast, autocratic leaders are either domineering or uninvolved and do not consult the opinions of others. Research suggests that the interaction of leadership style with the type of task and group is particularly relevant (e.g. Fiedler, 1993, cited in Goethals, 2005; see also Gastil, 1997), with democratic leadership being apparently more productive when experimental groups are given moderately or highly complex tasks, though the link between democratic leadership and satisfaction was found not to be particularly strong or uniform (Gastil, 1994). Gastil (1993a, cited in Gastil, 1993b) also identified a number of obstacles to small group democracy, including excessive meeting length, unequal levels of commitment and involvement of different group members, clique formation and mini-consensus (formed in and/or outside meetings), differences in communication skills and styles, and intense interpersonal conflicts.

Decision-making in an educational context

Observational data from educational contexts detected a number of the above-mentioned social factors in, for instance, awarding meetings and Angoff meetings (e.g. Murphy *et al*., 1995; Brennan and Lockwood, 1980). In these studies, dominant group members were found to unduly influence the consensus opinion; there was evidence of individuals being under pressure to conform when presented with a consensus opinion; the meetings were strongly influenced by decisions taken by the Chair or by the ways in which the Chair exercised his or her role, etc. Regarding democratic (non-hierarchical) vs. autocratic (hierarchical) processes in standardisation meetings, Baird *et al*. (2004) note that, according to the questionnaire responses they collected, examiners preferred having a hierarchical discussion to having no discussion in standardisation meetings, and there was some preference for non-hierarchical rather than hierarchical discussion.

Black and Curcin (*in submission*; see also Black, Suto and Bramley, *in submission*) investigated the relationship of various group dynamics factors in expert panels deciding on definitive marks on seeding items with subsequent marker agreement. The researchers coded the discussion surrounding the decisions regarding each mark in five OCR units in terms of level of contention (which encapsulates factors such as minority influence, conformity, cohesion) and democracy levels (subsumes leadership style), as well as discussion time, and investigated these “meeting features” in relation to levels of subsequent marker agreement with the definitive marks.

While democracy was found to be related to P_0 for only two of five units under investigation and further investigation was deemed necessary, the other two features (contention and discussion time) were strongly related to P_0 for all units (higher contention and longer discussion time were associated with lower agreement). Indeed, these two features were two of the strongest single predictors of marker agreement (with similar or higher levels of prediction as maximum mark or item type) and can be seen as an expression of many of the other features that affect marking task demands. Thus, the authors suggest that these meeting features might each be thought of as a composite of the interaction of question features, mark scheme features and response features and thus might be considered as useful heuristics for prediction of subsequent marker agreement.

Conclusion

The overview given here clearly leads to a conclusion that the more objective an item and consequently the more constrained the mark scheme, the higher level of marker agreement will be achieved, though this can become complicated by, for instance, the nature of candidate response. However, marking reliability is only one of the many concerns of assessment. As Newton (1996) points out, changing the format of questions or mark schemes to increase marker agreement may threaten assessment validity as, for instance, more constrained questions may fail to measure the desired construct in some subjects appropriately. On the other hand, low marking reliability also has a negative effect on validity as the same marks given by different markers cannot be assumed to mean the same thing. More detailed and integrated knowledge of various factors that affect marker agreement which can be gleaned from item-level investigations in the context of seeding, as well as from investigations of group dynamics in expert panels deciding on definitive marks, could equip awarding bodies with an understanding of the levels of marker agreement that could be expected in different contexts and that could realistically be aspired to. This in turn could perhaps help boost reliability by improving marker agreement prediction, monitoring, feedback and training practices, without the need for resorting to over-constrained questions in inappropriate contexts.

References

Asch, S.E. (1951). Effects of group pressure on the modification and distortion of judgements. In: H. Guetzkow (Ed.), *Groups, Leadership and Men*. 177–190. Pittsburgh: Carnegie.

Asch, S.E. (1956). Studies of independence and conformity: a minority of one against a unanimous majority. *Psychological Monographs*, **70**, 9, (whole no. 416), 1–70.

Baird, J-A., Greatorex, J. & Bell, J.F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education, Principles, Policies and Practices*, **11**, 3, 333–347.

Black, B. & Curcin, M. (*in submission*). Group dynamics in determining ‘gold standard’ marks for seeding items and subsequent marker agreement.

Black, B., Curcin, M. & Dhawan, V. (*in submission*). Investigating seeding items used for monitoring on-line marking: factors affecting marker agreement with the gold standard marks.

Black, B., Suto, W.M.I. & Bramley, T. (*in submission*). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement.

Bramley, T. (2008). *Mark scheme features associated with different levels of marker agreement*. Paper presented at the British Educational Research Association (BERA) annual conference, Heriot-Watt University, Edinburgh, September 2008.

Brennan, R.L. & Lockwood, R.E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement*, **4**, 219–240.

Briggs, D. (1980). A study of the influence of handwriting upon grades using examination scripts. *Educational Review*, **32**, 2, 185–193.

Briggs, D. (1970). The influence of handwriting on assessment. *Educational Research*, **13**, 1, 50–55.

Budescu, D.V., Rantilla, A.K., Yu, H.T. & Krelitz, T.M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behaviour and Human Decision Processes*, **90**, 1, 178–194. (cited in Kerr & Tindale, 2004).

Budescu, D.V. & Rantilla, A.K. (2000). Confidence in aggregation of expert opinions. *Acta Psychologica*, **104**, 371–98. (cited in Kerr & Tindale, 2004).

Bull, R. & Stevens, J. (1979). The effects of attractiveness of writer and penmanship on essay grades. *Journal of Occupational Psychology*, **52**, 53–59.

Cambridge Assessment (2009). The Cambridge Approach. Principles for designing, administering and evaluating assessment. Available at http://www.cambridgeassessment.org.uk/ca/digitalAssets/181348_cambridge_approach.pdf Accessed 15/02/10.

Crisp, V. (2007). *Do assessors pay attention to appropriate features of student work when making assessment judgements?* Paper presented at the International Association for Educational Assessment Annual Conference, Baku, September 2007.

Deutsch, M. & Gerard, H.B. (1955). A study of normative and informational influence upon individual judgement. *Journal of Abnormal and Social Psychology*, **51**, 629–636.

Fiedler, F.E. (1993). The leadership situation and the black box in contingency theories. In: M. M. Chemers, & R. Ayman (Eds.), *Leadership Theory and Research.* 1–28. San Diego, CA: Academic. (cited in Goethals, 2005).

Fitzpatrick, A. (1989). Social influences in standard setting: the effects of social interaction on group judgments. *Review of Educational Research*, **59**, 315–328.

Gastil, J. (1997). A Definition and Illustration of Democratic leadership. In: K. Grint (Ed.), *Leadership: Classical, Contemporary and Critical Approaches*. 155–178. Oxford: OUP.

Gastil, J. (1994). A meta-analytic review of productivity and satisfaction of democratic and autocratic leadership. *Small Group Research*, **25**, 3, 384–410.

Gastil. J. (1993a). *Meeting democracy: Participation and decision making in small groups*. Philadelphia: New Society Publishers.

Gastil, J. (1993b). Identifying obstacles to small group democracy. *Small Group Research*, **24**, 1, 5–27.

Goethals, G.R. (2005). Presidential leadership. *Annual Review of Psychology*, **56**, 545–570.

Greatorex, J. & Suto, W.M.I. (2006). *An empirical exploration of human judgement in the marking of school examination*s. Paper presented at the annual conference of the International Association for Educational Assessment, 21–26 May, Singapore.

Harvey, N., Harries, C. & Fischer, I. (2000). Using advice and assessing its quality. *Organizational Behaviour and Human Decision Processes*, **81**, 52–73. (cited in Kerr & Tindale, 2004).

Hudson, G., Donahue, B.H., Rutt, S. & Schagen, I. (2007). *Is electronic marking just about efficiency? Further analysis of electronic marking data to investigate factors related to marking reliability*. DRS Data Services Limited.

Janis, I. L. (1972). *Victims of Groupthink*. Boston: Houghton Mifflin. (cited in van Avermaet, 1988).

Kerr, N.L. & Tindale, R.S. (2004). Group performance and decision making. *Annual Review of Psychology*, **55**, 623–55.

Levine, J.M. & Moreland, R.L. (2006). *Small Groups*. New York and Hove: Psychology Press.

Lewin, K. & Lippitt, R. (1938). An Experimental Approach to the Study of Autocracy and Democracy: A Preliminary Note. *Sociometry*, **1**, 3/4, 292–300. (cited in Gastil, 1997).

Lewin, K., Lippitt, R. & White, R.K. (1939).Patterns of aggressive behaviour in experimentally created "Social Climates." *Journal of Social Psychology*, **10**, 271–279. (cited in Gastil, 1997).

Markham, L.R. (1976). Influences of Handwriting Quality on Teacher Evaluation of Written Work. *American Educational Research Journal*, **13**, 4, 277–283.

Massey, A. (1983). The effects of handwriting and other incidental variables on GCE 'A' level marks in English Literature. *Educational Review*, **35**, 1, 45–50.

Massey, A. & Foulkes, J. (1994). Audit of the 1993 KS3 Science national test pilot and the concept of quasi-reconciliation. *Evaluation and Research in Education*, **8**, 119–132.

Massey, A.J. & Raikes, N. (2006). *Item level examiner agreement*. Paper presented at the 2006 Annual Conference of the British Educational Research Association, 6–9 September 2006, University of Warwick, UK.

Moscovici, S. (1976). *Social Influence and Social Change*. London: Academic Press.

Moscovici, S. & Zavalloni, M. (1969). The group as the polarizer of attitudes. *Journal of Personality and Social Psychology*, **12**, 125–135.

Murphy, R.J.L. (1982). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology*, **52**, 1, 58–63.

Murphy, R.J.L. (1979). Removing the Marks from Examination Scripts before Re-Marking Them: Does It Make Any Difference? *British Journal of Educational Psychology*, **49**, 1, 73–78.

Murphy, R.J.L. (1978). Reliability of marking in eight GCE examinations. *British Journal of Educational Psychology*, **48**, 2, 196–200.

Murphy, R., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J. & Gower, R. (1995). *The dynamics of GCSE awarding (DOGA)*. Report of a project conducted for the School Curriculum and Assessment Authority, School of Education, University of Nottingham.

Myers, D.G. (1982). Polarizing effects of social interaction. In: H. Brandstätter, J. H. Davis & G. Stocker-Kreichgauer (Eds.), *Group Decision Making*. 125–157. New York: Academic Press. (cited in van Avermaet, 1988).

Newton, P.E. (1996). The Reliability of Marking of General Certificate of Secondary Education Scripts: Mathematics and English. *British Educational Research Journal*, **22**, 4, 405–420.

Ofqual (2009) *Code of practice for GCSE, GCE, and AEA*, April 2009.

Pinot de Moira, A., Massey, C., Baird, J. & Morrissy, M. (2002). Marking consistency over time. *Research in Education*, **67**, 79–87.

Phillips, K.W. & Lewin Loyd, D. (2006). When surface and deep-level diversity collide: The effects on dissenting group members. *Organizational Behaviour and Human Decision Processes*, **99**, 143–160.

Sniezek, J.A. (1992). Groups under uncertainty: an examination of confidence in group decision making. *Organizational behavior and human decision processes*, **62**, 159–174. (cited in Kerr & Tindale, 2004).

Suto, W.M.I. & Greatorex, J. (2008a). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*, **34**, 1, 1–21.

Suto, W.M.I. & Greatorex, J. (2008b). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policies and Practice*, **15**, 1, 73–89.

Suto, W.M.I. & Nádas, R. (2009) Why are some GCSE examination questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features. *Research Papers in Education*, **24**, 3, 335–377.

Suto, W.M.I. & Nádas, R. (2008). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education*, **23**, 4, 477–497.

Suto, W.M.I., Nádas, R. & Bell, J.F. (2009). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education*. (Published online to date).

van Avermaet, E. (1988). Social Influence in Small Groups. In: M. Hewstone, W. Stroebe, J-P. Codol & G. M. Stephenson (Eds.), *Introduction to Social Psychology*. 350–380. Oxford: Basil Blackwell.

Vidal Rodeiro, C. (2007). Agreement between outcomes from different double-marking models. *Research Matters: A Cambridge Assessment Publication*, **4**, 28–34.

Whetton, C. & Newton, P. (2002). *An evaluation of on-line marking*. Paper presented at the 28th International Association for Educational Assessment Conference, Hong Kong SAR, China, September.

Wilmut, J., Wood, R. & Murphy, R. (1996). *Review of Research into the Reliability of Examinations*. A discussion paper prepared for the School Curriculum and Assessment Authority.

Wilmut, J. (1984). A pilot study of the effects of complete or partial removal of marks and comments from scripts before re-marking them. *AEB Research Report RAC315*.

Zimbardo, P. & Leippe, M.R. (1991). *The Psychology of Attitude Change and Social Influence*. New York: McGraw Hill. (cited in Murphy *et al.*, 1995).

NEW TECHNOLOGIES

Why use computer-based assessment in education? A literature review

Matt Haigh Research Division

Introduction

The aim of this literature review is to examine the evidence around the claims made for the shift towards computer-based assessment (CBA) in educational settings. In this examination of the literature a number of unevidenced areas are uncovered, and the resulting discussion provides the basis for suggested further research alongside practical considerations for the application of CBA.

The review looks at academic literature from UK and international contexts, examining studies that are based in educational settings from primary education to higher education. It should be noted that the literature identified predominantly emerges from higher education contexts in the UK.

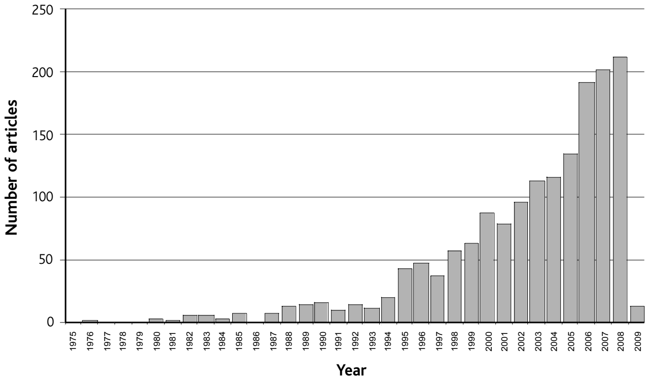
Background

CBA first emerged in educational settings in the 1950s and has undergone a steady expansion in use. Burkhardt and Pead (2003) provide a useful summary of the development of CBA in educational settings for each decade between 1950 and 2000:

> *1950s: Early computers offered games, puzzles and 'tests'; compilers were designed to identify errors of syntax, and later of style, in computer programs.*
>
> *1960s: The creators of learning machines, in which assessment always plays a big part, recognised the value of computers for delivering learning programmes.*
>
> *1970s: The huge growth of multiple-choice testing in US education enhanced the attractions of automatic marking, in a self-reinforcing cycle.*
>
> *1980s: A huge variety of educational software was developed to support learning, with less emphasis on assessment.*
>
> *1990s: Along with the continuing growth of multiple-choice testing, integrated learning systems, a more sophisticated development of the learning machines of the 1960s, began to be taken more seriously.*
>
> *Since the 1990s, the explosive growth of the internet has begun to raise the possibility that testing online, on-demand might replace the traditional 'examination day' model, although many technical and educational challenges remain.*
>
> (Burkhardt and Pead 2003, p.134)

This history highlights the varying degree to which assessment has formed part of technology-facilitated pedagogy, along with the dangers of allowing technology to dictate assessment practices such as with the permeation of multiple-choice testing in the US during the 1970s detailed by Clarke, Madaus, Horn, and Ramos (2000).

The accompanying expansion in research activity can be illustrated by interrogating online-databases and filtering by year of publication as illustrated in Figure 1. This indicates that CBA developments in the mid-1990s, highlighted in the quote above, spawned a dramatic increase in the research literature available.



Articles with keyword "Computer-Based Asessment" from SCOPUS social science database

Figure 1: An illustration of CBA research activity

Note that CBA covers a broad range of assessment types, from high-stakes multiple-choice tests through to compilation of assessment evidence in electronic portfolios. This review encompasses this range, however it is quite plausible that the research discussed may only apply to a subset of these assessment types and the reader should consider this caveat throughout.

Strategy for the literature review

In line with approaches to reviewing that make explicit the approach for searching and managing the literature, this section sets out a description of the approach taken. Initially the literature searched emerged from personal professional knowledge. This was then expanded via a number of strategies:

- The use of bibliographic databases and search engines (Scopus, British Education Index, ERIC, Web of Knowledge, Psycinfo, Zetoc, Google Scholar, Directory of Open Access Journals, Education-line, Educational Evidence Portal, Multiverse, Intute);
- Identification of a number of key journals: *British Journal of Educational Technology, Assessment in Education: Principles, Policies and Practice and ALT-J*; a subsequent search of these journal indexes provided additional literature.
- The citation index tools in SCOPUS were also used to identify the most frequent citations in the literature.

Given the number of items of literature emerging from these approaches and the scope of this article, no systematic attempt was made to reference additional repositories of 'grey literature'[1]. All emerging literature was collated and categorised using bibliographic software. The criteria for inclusion were:

- Research carried out in educational institutions and available in the public domain (this excluded work-based training and the use of CBA for recruitment).
- Research included a component of evaluation[2] of the use of CBA.
- A focus on research published post-1995: given the development of technology, particularly the explosion of internet use in the mid-1990s, older studies evaluated different computing technology; therefore pre-1995 studies have only been included as an exception.

It should be noted that a significant proportion of the literature identified was based on case-study methodologies.

Three approaches were taken to extract salient themes from the research:

- Using a set of key questions that were explored using the literature.
- Identifying literature which presents an overview of CBA use, and extracting key themes.
- Use of the tagging system[3] employed to code literature in the bibliographic software.

Figure 2 illustrates this approach.

Overview of the literature emerging from each strategy

1. Using key questions to identify themes

Why use Computer-Based Assessment in Education?

The most immediate claim that emerges from key texts is that CBA is a facilitator of formative assessment (Brown, Race and Bull, 1999). A discussion of the relationship between CBA and formative assessment would seem inevitable given the relentless interest in Black and Wiliam's work encompassed in their publication 'Inside the Black Box' (Black and Wiliam, 1998), aspects of which have made their way into UK Government educational policy (DCSF, 2008). Therefore a further question emerges:

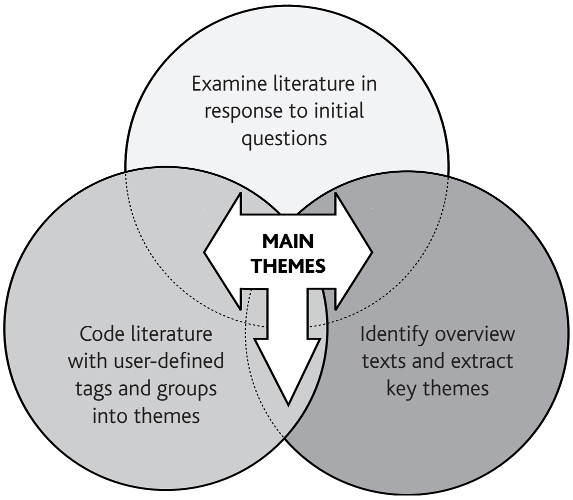

Examine literature in response to initial questions
MAIN THEMES
Code literature with user-defined tags and groups into themes
Identify overview texts and extract key themes

Figure 2: Illustration of the method used to identify themes

What is the relationship between CBA and Formative Assessment?

An examination of the recommended practice in Black and Wiliam's work does indicate areas of formative assessment practice on which CBA might have an impact, for example:

> *Feedback to any pupil should be about the particular qualities of his or her work, with advice on what he or she can do to improve, and should avoid comparisons with other pupils.…Tests and homework exercises can be an invaluable guide to learning, but the exercises must be clear and relevant to learning aims. The feedback on them should give each pupil guidance on how to improve, and each must be given opportunity and help to work at the improvement.*
>
> (Black and Wiliam, 1998, p.9)

CBA has the capability to provide feedback for each individual student and, with suitable mechanisms for analysing data, can provide feedback on each student's strengths and weaknesses in relation to their responses to assessment items.

The automated marking element of CBA has the potential to provide timely feedback to enable students to engage in self-assessment. However, feedback from CBA by itself is unlikely to develop the self-assessment skills of students, as Black and Wiliam point out:

> *For formative assessment to be productive, pupils should be trained in self assessment so that they can understand the main purposes of their learning and thereby grasp what they need to do to achieve.*
>
> (Black and Wiliam 1998, p. 10)

It is less clear how CBA may be used in relation to other points raised by Black and Wiliam such as ensuring "The dialogue between pupils and a teacher should be thoughtful, reflective, focused…" (p.12).

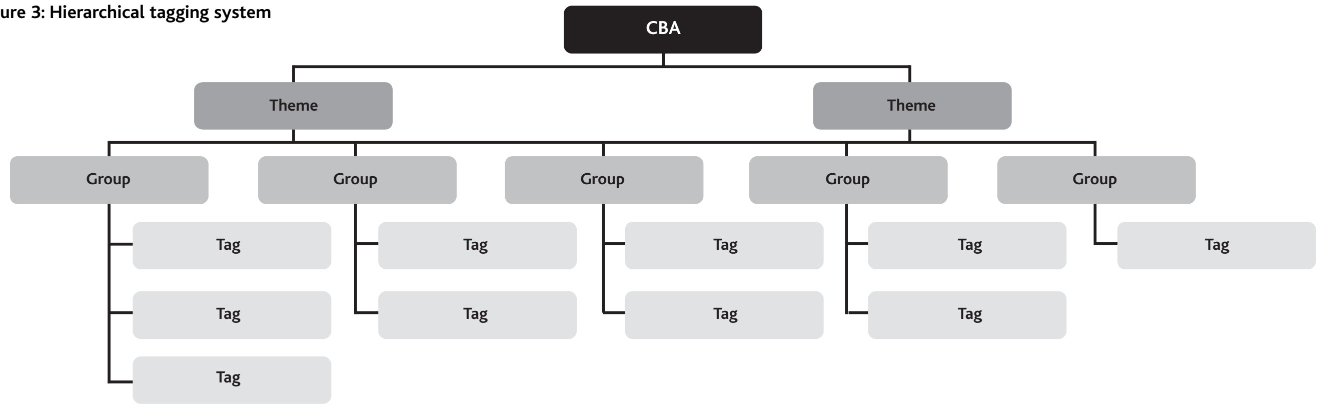
The next key question emerges from the discourse in 'Computer Assisted Education in Higher Education' (Brown *et al*., 1999), where the following statement is made:

[1] Documents not formally published through traditional channels, e.g. government technical reports, commercial product evaluations.

[2] Some papers, although set in the context of a CBA environment, were not evaluating the application of CBA per se, but often another aspect of the associated programme.

[3] In 'tagging' each piece of literature can be assigned any number of user-defined codes (e.g. 'higher education' 'Formative assessment' 'case-study') which are stored by the bibliographic software alongside the item in question. These tags can then be searched, for example, to find all literature with the 'higher-education' tag associated with it.

Figure 3: Hierarchical tagging system



> *…in most subject disciplines the use of information and communications technologies is expanding rapidly and students are learning a higher proportion of the curriculum using computer-based resources…The gap between how students learn and how they are assessed is widening.*
>
> (Brown *et al*., 1999, p.205)

This provokes the following line of inquiry:

What is the relationship between CBA and students' methods of learning?

The subject of the interrelationships between assessment and learning is much debated. This question will be considered by drawing on Gipps' theory of educational testing (Gipps, 1994), in which the relationship to learning is much discussed: "The implication of work in cognitive science for the assessment of student learning, is that we need to focus on the models that students construct for themselves" (p.29). Therefore, if the models employed by students in their learning are strongly built around a technology-supported environment, then there is a clear argument for the use of CBA in educational assessment.

Gipps also discusses the importance of a wider approach to assessment: "We need a much wider range of assessment strategies to assess a broader body of cognitive aspects than mere subject-matter acquisition." (p.10). The implication for CBA here is that if our 'broader body of cognitive aspects' includes those associated with technology use, then CBA would be the associated assessment strategy.

2. Other claims made from overview texts on the use of CBA

An examination of texts with an overview of CBA derives a number of further claims for CBA. First, there are those who advocate CBA for virtues of efficiency: both Brown *et al*. (1999) and Thelwall (2000) talk of reducing workload by automation; Bull and McKenna (2003) indicate that CBA can be used to decrease marking loads and ease administrative efficiency. In a similar vein Linn, Baker and Dunbar (1991) put forward eight criteria for the evaluation of new assessment types, one of which sits under the heading 'cost and efficiency'.

It is interesting that the notion of efficiency is entering the educational discourse; it could be proposed that this is a managerial function of CBA. However, it is possible to argue that education should be concerned with efficiency: Brown *et al*. (1999) talk of the reduction in resource per student in higher education and the difficulties in extending traditional assessment to meet demand. There is concern that the term 'efficiency' is being used as a cover for a reduction in quality of education, and a justification for the reduction in public-spending on education (Welch, 1998). This implies that efficiency is directly related to the quality of education, and it is on this basis that the relationship between CBA and efficiency can be an educational issue.

Also emerging from key texts is a theme of motivation: Bull and McKenna, (2003) propose that CBA allows one to increase frequency of assessment to motivate students to learn and encourage students to practice skills. This seems to imply that increased frequency of assessments is a factor in motivating students. This is in contrast to other research indicating that testing is seen to decrease students' motivation to learn (Harlen and Deakin Crick, 2003).

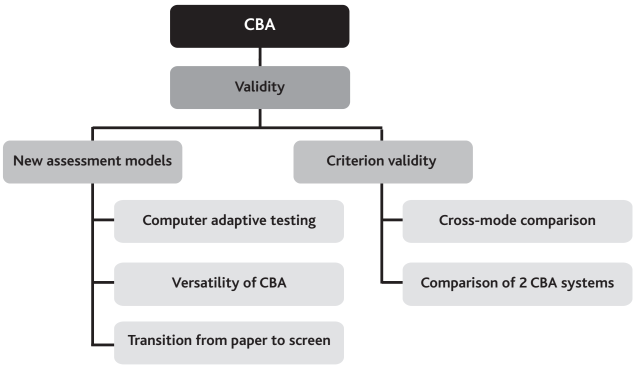
Both Thelwall (2000) and Conole and Warburton (2005) raise the issue of the difficulty of institutional implementation and wide-scale use associated with CBA, however factors that alleviate these difficulties (e.g. the development of staff knowledge of CBA) have also been proposed (Ely, 1999).

Themes emerging from tagging in bibliographic software

As the most relevant literature was collated, the content was coded with user defined tags in the bibliographic software. These codes could then be grouped to identify common elements, which were labelled 'groups'. In a similar process, these 'groups' were assembled into common elements called 'themes'. Figure 3 illustrates this hierarchical scheme of coding. The 'tags' with common concepts are collated into 'groups', which are further collated into themes.

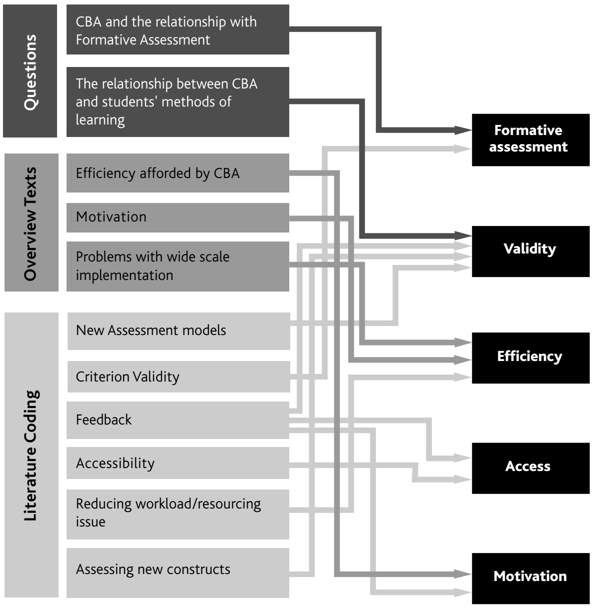
In all, 289 items of literature were examined; Figure 4 illustrates the application of the hierarchy to a set of tags. As an example, the tags 'computer adaptive testing', 'versatility of CBA' and 'transition from paper to screen' have all been put into a group labelled 'new assessment models'. The groups 'new assessment models' and 'criterion validity' have been put together under the theme 'validity'.

Figure 4: Application of the coding hierarchy



Syntheses of key themes identified in the literature

Using the three strategies (key questions, overview texts and tagging), the analysis of claims made for CBA provides a set of themes that can be seen to converge. The convergence is shown graphically in Figure 5 with the three strategies forming the first column. The second column indicates the key concepts arising from the first two strategies, and a set of groups arising from the tagging strategy. The arrows then show the links to five core themes that emerge across all three strategies.



Questions
CBA and the relationship with Formative Assessment
The relationship between CBA and students' methods of learning
Overview Texts
Efficiency afforded by CBA
Motivation
Problems with wide scale implementation
Literature Coding
New Assessment models
Criterion Validity
Feedback
Accessibility
Reducing workload/resourcing issue
Assessing new constructs
Formative assessment
Validity
Efficiency
Access
Motivation

Figure 5: Representation of the emergence of key themes

These five emergent key themes in the existing literature can be more accurately specified as follows:

- Use of CBA to improve the efficiency of assessment programmes (efficiency).
- Use of CBA to facilitate or enhance formative assessment practices (formative assessment).
- The effect of CBA on the validity of assessments (validity).
- Use of CBA to facilitate access to assessments (access).
- The effect of CBA on student motivation (motivation).

It is worth noting at this stage that the themes of efficiency and, to a certain extent, validity indicate a system-centric view of education (pre-occupied with measurement and effective use of resources). The themes of formative assessment, access and motivation indicate a more learner-centred view of education.

Examining evidence in the literature

Evidence for efficiency

It would appear that very little empirical evidence exists that CBA improves efficiency. Loewenberger and Bull (2003) struggled to reach conclusions on the cost-effectiveness aspect of efficiency, but hypothesised that CBA would be more suitable for larger groups. Their report indicates that due to factors associated with immaturity of ICT use and resistance to change:

> *…it becomes extremely difficult to obtain hard data that conclusively demonstrate the cost-effectiveness of CBA. Recommendations and hypotheses in this study could form the basis for further research. It would seem that the evidence for efficiency is far from well-developed.*
>
> (Loewenberger and Bull, 2003, p.38)

Evidence for CBA facilitating formative assessment

The claims made for CBA facilitating formative assessment are largely derived from a number of evaluations of case studies, so generalisations are difficult to establish unless most cases have the same outcome. In order to provide a framework for exploring the studies relating to formative assessment, the emerging research can be referenced to The Assessment Reform Group's 10 principles for formative assessment:

1. *Assessment for learning should be part of effective planning of teaching and learning.*
2. *Assessment for learning should focus on how students learn.*
3. *Assessment for learning should be recognised as central to classroom practice.*
4. *Assessment for learning should be regarded as a key professional skill for teachers.*
5. *Assessment for learning should be sensitive and constructive because any assessment has an emotional impact.*
6. *Assessment should take account of the importance of learner motivation.*
7. *Assessment for learning should promote commitment to learning goals and a shared understanding of the criteria by which they are assessed.*
8. *Learners should receive constructive guidance about how to improve.*
9. *Assessment for learning develops learners' capacity for self-assessment so that they can become reflective and self-managing.*
10. *Assessment for learning should recognise the full range of achievements of all learners.*

(Assessment Reform Group, 2002, p.2)

A number of studies cite the availability of immediate feedback for students as a key benefit in this area (Ashton and Wood, 2006; Bull, Quigley and Mabbott, 2006; Peat and Franklin, 2002). This relates to other research that indicates immediacy of feedback is important in the self-assessment process; however, factors other than immediacy are also shown to be important in Clariana, Ross and Morrison (1991). Despite this, Topping, Samuels and Paul (2007) make a strong case for the educational benefits of timely feedback. This in turn relates to the strand of formative assessment related to self-assessment (principle no.9 above).

Studies also indicated that CBA was able to shed more light on student's difficulties with subject knowledge (Jean, Delozanne, Jacoboni, and Grugeon, 1998) or identify students' methods of learning (Bull *et al.*, 2006). This links to the ideas of students receiving constructive guidance (principle no.7 above).

Studies such as Hunt, Hughes, and Rowe (2002) and Lowry (2005) make claims that improved student performance was related to the formative use of CBA. However, it is difficult to establish attainment gains as a direct result of the use of CBA as the meta-analysis of 23 studies by Fuchs and Fuchs (1986) reported significant attainment gains by those involved in non-CBA formative-assessment based interventions.

The use of CBA was reported to increase dialogue between student and teacher in two case studies (McGuire, 2005; Nicol, 2007), this chimes strongly with one of the key points made by Black and Wiliam (1998):

> *The dialogue between pupils and a teacher should be thoughtful, reflective, focused to evoke and explore understanding, and conducted so that all pupils have an opportunity to think and to express their ideas.*
>
> (Black and Wiliam, 1998, p.12)

Therefore evidence that indicates the use of CBA encourages dialogue is starting to align with the ideals of formative assessment practice.

On the other side of the coin, the studies of the formative use of CBA were also scoured for evidence of any negative impacts associated with their implementations. One recurring theme was the difficulty in demonstrating equivalence between CBA and paper-based formats (Ashton and Wood, 2006; Johnson and Green, 2004).

Another negative impact of CBA was the amount of time taken for both students and teachers to 'learn the system' (Jean *et al*., 1998; McGuire, 2005). This has implications in terms of large scale implementations of CBA as illustrated by Nicol (2007) and the difficulty of institutional implementation and wide-scale use highlighted by Conole and Warburton (2005).

There was also evidence that particular systems were promoting a mechanistic approach and confined to the assessment of lower-order skills (McKenna, 2001). However, other studies such as Ridgway and McCusker (2003) contradict this by implying that CBA facilitates the assessment of higher order skills. Together the studies indicate that the assessment of lower or higher order skills may not be a function of CBA, but the way in which it is used.

It would seem that the key themes emerging from the review of studies linking CBA and formative assessment, in order of prevalence and sufficiency of evidence, are as follows:

- The use of CBA for instant feedback and self-assessment.
- The use of CBA to facilitate anytime-anywhere access to formative assessment.
- Concerns regarding equivalence with paper-based assessments.
- Time taken for students to familiarise with the computer interface.

If we return to the framework of the 10 principles at the beginning of this section, the themes identified in the research on formative assessment focus very strongly on the use of CBA to uphold principle number 9 – opportunities for self-assessment. This is well evidenced in the available research and is exemplified by the findings in Bull *et al*. (2006) and McGuire (2005).

However, this leaves any claim that CBA can enhance a number of the principles for formative assessment un-evidenced from the research. Only one study demonstrated the use of CBA in focusing how students learn (Peat and Franklin, 2002). Similarly, there was only limited evidence on how CBA helped place AfL as central to classroom practice (McGuire, 2005). Little research evidence is available on the use of CBA to support the remaining principles.

In examining these studies related to the formative use of CBA, the literature indicates that only one key aspect of formative assessment is significantly evidenced, namely the capacity of CBA for instant feedback and providing opportunities for self-assessment or reflection. It is clear that there is much less evidence available for how CBA supports the remaining principles outlined in this section.

Evidence for CBA improving the validity of assessments

The American Psychological Association (APA) 'Guidelines for Computer based Tests and Interpretations' state that "the validation of computer-based tests and protocols does not differ in kind from the validation of tests generally" (APA, 1986, p.19).

It is worth emphasising the importance given to validity in evaluating new forms of assessment:

> *The arguments, pro and con, regarding traditional and alternative forms of assessment need to give primacy to evolving conceptions of validity if, in the long run, they are to contribute to the fundamental purpose of measurement – the improvement of instruction and learning.*
>
> (Linn *et al*., 1991, p.20)

Russell, Goldberg and O'Connor (2003) provide a useful summary of some aspects of validity research since 1986 which cites evidence on the following areas:

- The inability to review or revise responses (this, in particular, is a feature of computer adaptive testing) has a negative effect on examinee performance.
- Graphical display issues, such a screen size and resolution, affect examinee performance.
- Familiarity with computers plays a role in test performance.

These three areas refer to Messick's (1989) concept of 'construct irrelevant variance' which becomes a recurring theme in the reviewed literature. These points also serve as a useful illustration of three very different sources of construct irrelevant variance:

- The method by which assessment items are sequenced.
- Aspects of screen display.
- The characteristics of the student in relation to ICT.

Sources of construct irrelevant variance in an on-screen assessment of ICT skills are also explored by Threlfall, Nelson and Walker (2007), who approach the analysis by examining 'sources of difficulty' and then identifying those that are linked to the construct, and those that are irrelevant to the construct.

One recurrent feature of CBA research are studies designed to yield comparisons with 'equivalent' paper-based tests – this is evaluating the traditional dimension of criterion-related concurrent validity (often referred to as cross-modal validity in the literature). A meta-analysis of such studies by Bunderson *et al*. (1989) demonstrated better performance in computerised tests in 3 cases; no difference in 11 cases; and better performance in paper tests in 9 cases. The meta-analysis revealed some potential reasons for the modal differences:

- Aspects of item delivery.
- Aspects of item presentation.
- The students' background characteristics – particularly in relation to ICT.

Note that these points concord with the findings by Russell *et al*. (2003) above. Some research has focussed on the students' characteristics:

> *In summary, establishing a model that fully accounts for test performance differences may be some time away, however it seems*

> *critical at this time to further this line of research. Based on our review and these results, we anticipate that computer familiarity is the most fundamental key factor in the test mode effect.*
>
> (Clariana and Wallace, 2002, p.601)

Huff and Sireci, (2001) examine issues regarding validity in computer-based testing. First they look at the evidence used in favour of CBA to enhance validity and conclude that most of these arguments centre on: 1) increasing construct representation, and 2) improving measurement precision.

They go on to state that the claims that computer-based testing can enhance validity can be traced to at least four current developments:

- Innovative item formats.
- Computerised-adaptive testing technology.
- Cognitively principled CBT design.
- Automated scoring.

In the same article, perceived threats to validity are also explored, namely construct under-representation and the introduction of construct irrelevant variance in CBT tests.

Regarding 'construct under-representation', some argue that CBA can improve the construct validity of a test in the case of assessments of problem solving skills (Ridgway and McCusker, 2003) and students' cognitive strategies (Nunes, Nunes and Davis, 2003). Some attempts have been made to provide a more empirical demonstration of construct validity in assessments of students' cognitive strategies (Wirth and Klieme, 2003).

From the discussion above, it is clear that there are many areas of 'construct irrelevant variance' to explore. Some inroads have been made with regards to identifying student background factors such as familiarity with ICT that have effects, but there clearly remain many areas of research activity left to explore in this area, particularly as CBA continues to evolve.

It appears that CBA introduces new sources of 'construct irrelevant variance' that, unmitigated, may reduce the validity of assessments. On the plus side, there is now a growing body of evidence indicating that CBA can facilitate the assessment of new constructs such as problem solving and meta-cognition.

Evidence for CBA facilitating access to educational assessments

The literature in this area can be divided into two further categories:

- The use of CBA to facilitate accessibility to assessments for individuals with disabilities.
- The use of CBA to facilitate access to assessments on-demand.

Taking the former, it is suggested that the increasing use of computer-based aids for those with disabilities make CBA an easier form of assessment to take advantage of these. For example, Bennett (1999) states:

> *From the perspective of task comparability, CBT offers substantial promise. One reason is that computers have become life-style accommodations for people with disabilities…an industry has evolved that produces dozens of alternative devices for getting information into and out of a personal computer.*
>
> (Bennett, 1999, p.181)

Empirical research is not evident in this area, which may be down to the small numbers of students with disabilities taking part in large-scale computer-based assessments. There is a warning associated with the use of features to enable accessibility, that un-checked they could evolve into threats to validity, that is, providing unfair assistance to particular students (Hansen, Mislevy, Steinberg, Lee and Forer, 2005) .

The latter aspect of accessibility, that CBA facilitates on-demand testing, can be illustrated using case studies. For example, an evaluation of CBA in undergraduate level Chemistry (Lowry, 2005) collected some qualitative feedback from students who indicated one benefit of CBA was the usefulness of being able to access the material at any time. The growth of the internet has clearly offered opportunities of online testing on demand:

> *Online students are able to take advantage of the accessibility of online assessment tasks from a variety of locations. They may receive valuable 'just in time' feedback from their teachers in order to make meaningful, timely decisions and judgements about their own learning.*
>
> (Northcote, 2002, p.623)

It would appear that there are further areas of empirical work to be done with regard to CBA improving accessibility to assessments for those with disabilities, particularly as increasing numbers participate in CBA programmes.

It would appear that commentators believe CBA has clear benefits in offering accessibility to those with disabilities. However, the lack of empirical research means that evidence is still awaited.

Evidence for CBA effects on student motivation

Harlen and Deakin Crick (2003) provide a useful framework for examining motivation through their meta-analysis of 19 studies linking motivation and testing. However, none of the studies examined involved the application of CBA. Even if it were assumed that findings would transfer to the CBA environment, they largely imply a negative association between testing and motivation. However, the hypotheses put forward by, for example, Bull and McKenna, (2003) and McKenna (2001) were that CBA improves student motivation. It has been difficult to find much evidence of motivational effects specifically associated with CBA, this is clearly an area that is ripe for further research.

Conclusions

The literature review has identified 5 themes associated with the evaluation of CBA.

In none of these areas was there comprehensive empirical evidence in the existing literature to back up the claims made for CBA.

Much of the evidence has emerged from case-study methodologies (particularly in the area of formative assessment), meaning that opportunities for generalisation are limited.

The two strongest themes are those of validity, which has been considered in a number of contexts, and formative assessment, which has evidence compiled from a number of case studies.

When the evidence gathered regarding the use of CBA in all five areas is scrutinised, a number of areas for further exploration and research activity emerge:

- Evidence primarily emerges for Higher Education contexts, suggesting that more work could be done to identify issues specifically related to secondary or primary levels of education.

- There is a possible need to evaluate cost effectiveness in a more conclusive manner.
- It would be more comprehensive to evaluate CBA against the additional criteria for formative assessment from the Assessment Reform Group that were discussed in the evaluation of evidence for the formative use of CBA.
- Aspects of validity relating to construct-irrelevant variance could be explored in the context of computer-based tests.
- There is a need to provide more empirical evidence on the impact that CBA has in supporting access.
- Work could be undertaken to identify if there is a link between the use of CBA and student motivation.
- A critical and structured review of the 'grey-literature' about CBA.

It is also possible to derive a number of practical aspects from the literature reviewed here, which will be of use to those considering how CBA may or may not improve the assessment experience at their learning institution:

- CBA does not have a strong empirical basis for efficiency claims – therefore the literature would suggest caution if the prime motivation for the introduction of CBA is efficiency.
- CBA has a strong case for improving self-assessment opportunities (particularly in the case of Higher Education).
- CBA has a limited evidence base for facilitating full formative assessment practice – therefore the introduction of CBA alone is unlikely to lead to full scale adoption of formative assessment.
- The literature indicates that CBA does have the opportunity to facilitate access to educational assessments.
- There is weak evidence for the motivational effects of CBA.

These practical pointers, although primarily of interest to those currently considering the use of CBA in education, will be important considerations for researchers undertaking empirical work about CBA in the future.

References

APA (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC.: American Psychological Association.

Ashton, H. & Wood, C. (2006). Use of Online Assessment to Enhance Teaching and Learning: the PASS-IT Project. *European Educational Research Journal*, **5**, 2, 9.

Assessment Reform Group (2002). *Assessment for learning: 10 principles*. Assessment Reform Group. Retrieved February 26, 2009, from http://www.assessment-reform-group.org/CIE3.PDF.

Bennett, R.E. (1999). Computer-based testing for examinees with disabilities: On the road to generalized accommodation. *Assessment in higher education: Issues of access, quality, student development, and public policy*, 181–191.

Black, P. & Wiliam, D. (1998). *Inside the Black Box: Raising Standards Through Classroom Assessment*. School of Education, King's College London.

Brown, S., Race, P. & Bull, J. (1999). *Computer-assisted Assessment in Higher Education*. London: Kogan Page.

Bull, J. & McKenna, C. (2003). *Blueprint for Computer-assisted Assessment.* 1st ed. London: Routledge.

Bull, S., Quigley, S. & Mabbott, A. (2006). Computer-based formative assessment to promote reflection and learner autonomy. *Engineering Education*, **1**, 1, 8–18.

Burkhardt, H. & Pead, D. (2003). Computer-based assessment: a platform for better tests? *Whither Assessment*, 133–148. London: Qualifications and Curriculum Authority.

Bunderson, C.V., Inouye, D.K. & Olsen, J.B. (1989) The four generations of computerized educational measurement. In: R.L. Linn (Ed.). *Educational measurement*. 3rd ed., 13–103. New York: American Council on Education.

Clariana, R. & Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, **33**, 5, 593–602.

Clariana, R., Ross, S. & Morrison, G. (1991). The effects of different feedback strategies using computer-administered multiple-choice questions as instruction. *Educational Technology Research and Development*, **39**, 2, 5–17.

Clarke, M.M., Madaus, G.F., Horn, C.L. & Ramos, M.A. (2000). Retrospective on educational testing and assessment in the 20th century. *Journal of Curriculum Studies*, **32**, 159–181.

Conole, G. & Warburton, B. (2005). A review of computer-assisted assessment. *ALT-J Research in Learning Technology*, **13**, 1, 17–31.

DCSF. (2008). *The Assessment for Learning Strategy*. Retrieved February 25, 2009, from http://publications.teachernet.gov.uk/default.aspx?PageFunction=productdetails&PageMode=publications&ProductId=DCSF-00341-2008.

Ely, D.P. (1999). Conditions that facilitate the implementation of educational technology innovations. *Educational Technology*, **39**, 6, 23–27.

Fuchs, L.S. & Fuchs, D. (1986). Effects of systematic formative evaluation: a meta-analysis. *Exceptional Children*, **53**, 3, 199–208.

Gipps, C.V. (1994). *Beyond Testing: Towards a Theory of Educational Assessment*. London: Routledge.

Hansen, E.G., Mislevy, R.J., Steinberg, L.S., Lee, M.J. & Forer, D.C. (2005). Accessibility of tests for individuals with disabilities within a validity framework. *System*, **33**, 1, 107–133.

Harlen, W. & Deakin Crick R. (2003). Testing and Motivation for Learning. *Assessment in Education: Principles, Policy and Practice*, **10**, 2, 169–207.

Huff, K.L. & Sireci, S.G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice*, **20**, 3, 16–25.

Hunt, N., Hughes, J. & Rowe, G. (2002). Formative Automated Computer Testing (FACT). *British Journal of Educational Technology*, **33**, 5, 525–535.

Jean, S., Delozanne, E., Jacoboni, P. & Grugeon, B. (1998). Cognitive Profiles in Elementary Algebra: the PÉPITE Test Interface. *Education and Information Technologies*, **3**, 3, 291–305.

Johnson, M. & Green, S. (2004). *On-line assessment: the impact of mode on student performance*. Paper presented at the British Educational Research Association Annual Conference, Manchester, September 2004.

Linn, R.L., Baker, E.L. & Dunbar, S.B. (1991). Complex, Performance-Based Assessment: Expectations and Validation Criteria. *Educational Researcher*, **20**, 8, 15–21.

Loewenberger P. & Bull J. (2003). Cost-effectiveness analysis of computer-based assessment. *Alt-J – Association for Learning Technology Journal*, **11**, 2, 23–45.

Lowry, R. (2005). Computer aided self assessment–an effective tool. *Chemistry Education Research and Practice*, **6**, 4, 198–203.

McGuire, L. (2005). Assessment using new technology. *Innovations in Education & Teaching International*, **42**, 3, 265–276.

McKenna, C. (2001). Introducing computers into the assessment process: what is the impact upon academic practice? *Higher Education Close Up Conference*, **2**.

Messick, S. (1989). Validity. In: R.L. Linn (Ed.) *Educational measurement*. 3rd ed., 13–103. New York: American Council on Education.

Nicol, D. (2007). Laying a foundation for lifelong learning: Case studies of e-assessment in large 1st-year classes. *British Journal of Educational Technology*, **38**, 4, 668–678.

Northcote, M. (2002). Online assessment: foe or fix? *British Journal of Educational Technology*, **33**, 5, 623–625.

Nunes, C.A.A., Nunes, M.M.R. & Davis, C. (2003). Assessing the inaccessible: metacognition and attitudes. *Assessment in Education: Principles, Policy and Practice*, **10**, 3, 375–388.

Peat, M. & Franklin, S. (2002). Supporting student learning: the use of computer-based formative assessment modules. *British Journal of Educational Technology*, **33**, 5, 515–523.

Ridgway, J. & McCusker, S. (2003). Using computers to assess new educational goals. *Assessment in Education*, **10**, 3, 309–328.

Russell, M., Goldberg, A. & O'Connor, K. (2003). Computer-based testing and validity: A look back into the future. *Assessment in Education: Principles, Policy & Practice*, **10**, 3, 279–293.

Thelwall, M. (2000). Computer-based assessment: A versatile educational tool. *Computers and Education*, **34**, 1, 37–49.

Threlfall, J., Nelson, N. & Walker, A. (2007). *Report to QCA on an investigation of the construct relevance of sources of difficulty in the Key Stage 3 ICT tests*. Retrieved February 26, 2009, from http://www.naa.org.uk/libraryAssets/media/Leeds_University_research_report.pdf.

Topping, K.J., Samuels, J. & Paul, T. (2007). Computerized assessment of independent reading: Effects of implementation quality on achievement gain. *School Effectiveness and School Improvement*, **18**, 2, 191–208.

Welch, A.R. (1998). The cult of efficiency in education: comparative reflections on the reality and the rhetoric. *Comparative Education*, **34**, 2, 157–175.

Wirth, J. & Klieme, E. (2003). Computer-based assessment of problem solving competence. *Assessment in Education: Principles, Policy and Practice*, **10**, 3, 329–345.

RESEARCH METHODS

Is CRAS a suitable tool for comparing specification demands from vocational qualifications?

Jackie Greatorex and Nicky Rushton Research Division

Introduction

Historically, unitary awarding bodies and the national regulator[1] monitored standards of qualifications between awarding bodies, over time and between cognate qualifications at the same level, and this work continues. A key reason for conducting such work is to avoid inequalities and inequities which would be created by the existence of easier routes to access further study or jobs.

Ideally standards are compared in terms of candidates' performance and in terms of the demands of the qualifications. When comparing new qualifications there is sometimes a lack of performance evidence[2] or assessment tasks[3] to form a robust sample from which generalisable research results can be drawn. In such cases comparability studies could focus on specifications[4] and the associated demands. However, studies restricted to one aspect of comparability (whether it be performance or demands) are limited.

One approach to comparing demands of qualifications is for experts to rate them on a scale of cognitive demands known as CRAS. CRAS was developed using academic qualifications. An issue deriving from its provenance may be that CRAS is not suitable for use with vocational qualifications which are different in nature and purpose to academic qualifications. Generally there are far more comparability studies about academic qualifications than VQ/VRQs[5]. In the present study we investigate whether CRAS is suitable for use in comparability studies which include VQs/VRQs.

Demands and difficulty

There is sometimes a lack of clarity about definitions of demands and difficulty.

In this article:

> *Task demands* refer to the actions (usually cognitive) a task is intended to require of typical members of the target group of learners. For example, candidates might be required to recall familiar information. Task demands generally relate to individual summative assessment tasks such as examination items. But task demands could also be related to an individual classroom activity or similar.
>
> *Specification demands* refer to the actions the specification is intended to require of typical members of the target group of learners in four areas: cognitive, affective, psychomotor and interpersonal. These specification demands might be explicit in the specification or they might be an underpinning ethos. For example, candidates might be required to recall information about a topic, empathise with another person's understanding of the topic, evaluate the other person's understanding to know what extra information they need and explain the relevant information to the

[1] Currently the national regulator of the awarding bodies is Ofqual.

[2] *Performance evidence* refers to students' work in the form of essays, artefacts, paintings, multiple choice responses and so on.

[3] *Assessment tasks* refers to examination questions, assignments, briefs for work-based projects and so on.

[4] The specification is: *The complete description – including optional and mandatory aspects – of the content, assessment arrangements and performance requirements for a qualification. A subject specification forms the basis of a course leading to an award or certificate. Formerly known as a 'syllabus'.* QCDA (undated)

[5] VQ refers to vocational qualifications and VRQ to vocationally related qualifications. These are very broad categories. Many vocational qualifications in England are NVQs (National Vocational Qualifications which: *are designed to recognise a candidate's competence in the workplace. They provide a statement to employers of skill, competence and knowledge in a particular sector.* (OCR, 2009a).
Vocationally related qualifications generally focus on an occupation or occupational sector: *Vocationally-Related Certificates enhance knowledge and build upon candidates' skills in preparation for a job*. (OCR, 2009b).

other person in an accessible manner. These examples of specification demands are cognitive and interpersonal. Specification demands relate to the specification; they are not about individual summative assessment tasks such as examination items.

Demanding refers to the extent to which a task or specification is intended to be challenging for typical members of the target group of learners.

Difficulty refers to "an empirical measure of how successful a group of students were on a question." (Pollitt *et al*., 2007, p.169). Relative difficulty can be measured as facility values; that is, the proportion of candidates giving the correct response to an item (Kline, 1986).

The notion of *intention* is crucial in clearly defining the concepts of task demands and difficulty. Task demands are about what typical members of the target group of learners are expected or intended to do to carry out a task. Difficulty is focused on the students' actual performance. Bloom (1956) emphasises this difference between what is intended and what actually happens in his work to develop a taxonomy of educational objectives.

The definitions of demands and difficulty used in this article are given above. However, there are various definitions of demand(s) which are used by other researchers for different contexts and purposes, for examples see Pollitt *et al*. (2007) or Barry (1997).

Awarding bodies and the national regulator have used various methods to compare the demands of academic qualifications. One approach has been to develop a questionnaire about task and specification demands, which senior examiners use to rate the task and specification demands of the various qualifications, for example, see Edwards and Adams (2003).

CRAS

Another approach to comparing task demands is to rate examination items on a scale of cognitive demands known as CRAS. The five aspects of the CRAS frame of reference given below are taken from Pollitt *et al*. (2007).

- "*Complexity:* The number of components or operations or ideas and the links between them." For example, using a single idea is less demanding than synthesising several ideas.
- "*Resources:* The use of data and information." For example, using all and only the information provided is less demanding than selecting the appropriate data.
- "*Abstractness:* The extent to which the student deals with ideas rather than concrete objects or phenomena." For example, work which deals with concrete objects is less demanding than mostly abstract work.
- "*Task strategy:* The extent to which the student devises (or selects) and maintains a strategy for tackling the question." For example, when a strategy is provided this is less demanding than when a strategy needs to be devised by the student.
- "*Response strategy:* The extent to which students have to organise their response." For example, giving the student a small number of possible responses to choose between is less demanding than them having to organise their own response.

The text in quotation marks is from Pollitt *et al*. (2007, p.186).

However, various concerns have been raised about the use of CRAS:

1. It has recently been used by QCA[6] to rate whole examinations rather than individual tasks; it was designed for the latter not the former.
2. In the context of comparing VQ/VRQs CRAS may not be suitable as it was developed using academic qualifications (Hughes *et al.*, 1998), which can be different in nature and purpose.
3. Whilst it can be used to compare task demands from academic qualifications it may not be applicable to VQ/VRQs specification demands.

In the present investigation the second and third concerns are addressed.

Crisp and Novaković (2009) used CRAS to compare the task demands from different centres for college-assessed units in a VRQ. They found that complexity, resources, task strategy and response strategy could be used to compare the task demands of various vocational assessments in one domain. However, abstractness was of less relevance.

In the present study we investigated whether CRAS was suitable for use in comparability studies about the assessment tasks and the specification of VQs/VRQs. To do this the CRAS frame of reference was compared with the frames of reference used in previous studies that compared the task and/or specification demands of VQ/VRQs.

Data

Data for the present study were taken from a series of comparability studies by awarding bodies or the national regulator about VQ/VRQs which are in the public domain (SCAA, 1995; Coles and Matthews, 1995, 1998; Arlett, 2002, 2003; Guthrie, 2003; QCA 2006, a and b). The data were the frames of reference used to compare qualifications in various studies about VQs/VRQs. The studies are outlined below.

Arlett (2002, 2003) and Guthrie (2003) used a modified version of Kelly's Repertory Grid to elicit the similarities and differences between VCE[7] qualifications from different awarding bodies in terms of summative assessment and specification requirements. The similarities and differences were used to develop items for a questionnaire on which senior examiners rated the various specifications, assessments, mark schemes or equivalent, and teacher support materials. The ratings were used to compare the qualifications. This approach was used in two vocational subjects.

SCAA (1995) asked subject experts to judge specifications, guidance to centres, examination papers and internal assessment[8] material/instructions and guidance against a series of factors drawn from the GNVQ[9] grading criteria and an UCLES[10] specification. The factors were:

> *Content: depth, breadth. Skills: factual recall, understanding and explanation, planning, investigation, analysis and evaluation, transferability (including the extent to which the student is encouraged to be adaptable and versatile) and application of skills*.
> (SCAA, 1995, p.4).

Breadth and depth refer to the breadth and depth of the qualification content which was studied and tested. The experts were also asked to judge whether the time requirements of the specification were likely to be met.

[6] The Qualifications and Curriculum Authority (QCA) was once the regulator of the awarding bodies. It was a predecessor of the Qualifications and Curriculum Development Agency (QCDA) and Ofqual.

[7] VCE or Vocational Certificate of Education is also sometimes referred to as the AVCE Advanced Vocational Certificate of Education. It was intended to replace the advanced General National Vocational Qualification (see below). In September 2005 VCEs were renamed GCE A-Levels (General Certificates of Education) in applied subjects. The specifications aim to give a broad introduction to vocational domains and to facilitate learning, teaching and assessment in work-related contexts. This information is from the Learning and Skills Council (2009).

QCA (2006a) used the following for subject experts to rate the level of cognitive demands of various multiple choice tests:

1. *Simple fact recall OR simple logic OR complex recall made easy by options*
2. *Complex recall including definitions*
3. *Showing understanding of a meaning; simple options, OR complex recall made difficult by options*
4. *Show understanding of a meaning: complex options*
5. *Apply reasoning with knowledge OR show understanding made difficult by options*. (QCA, 2006a, p.43).

A similar method and the same definitions of each level of cognitive demands were used in QCA (2006b) a comparability study of assessment practice for Door Supervision qualifications.

Additionally, in QCA (2006a) subject experts rated the plausibility of options in multiple choice tests. The reading difficulty of tests was identified and the accessibility of the questions was quantified by noting instances when important text was highlighted, perhaps by making it bold or italic. These issues, whilst they are not labelled "cognitive demand" by QCA (2006a), are similar to some of the items in Arlett (2002, 2003) and Guthrie (2003).

Coles and Matthews (1995, 1998) undertook a complicated methodology to qualitatively compare qualification learning outcomes, aims and content with a frame of reference, rather than compare the qualifications with one another. To create such a measure they adapted Bloom's taxonomy by adding a skills component based on the work of Gagne (1985) and Mitchel and Bartram (1994). Coles and Matthews (1995, 1998) argue that they needed the latter works to ensure that Bloom's taxonomy was not biased towards academic qualifications. Their frame of reference was based around recall, practical capability, interpretation, application, analysis and synthesis. They defined each term for the purposes of their study, then used this new frame of reference to classify the qualification and assessment requirements and to describe the specifications. Once the specification, learning outcomes and aims were classified in terms of the frame of reference the qualifications could be compared in detail.

In summary, the following were used as data in our study:

- The questionnaire items from Arlett (2002, 2003) and Guthrie (2003).
- SCAA's criteria, as well as the issue of time.
- QCA's levels of cognitive demands, plausibility of multiple choice options, reading difficulty and accessibility of text.
- Coles' and Matthews' (1995, 1998) frame of reference.

Procedure

The authors classified the data into three groups:

1. Referring to one of the five aspects of CRAS.
2. 'Other' (referring to task and/or specification demands not covered by CRAS).
3. 'Not' (referring to something which was not task and/or specification demands).

Initially one researcher classified the data. The judgements were checked by a second researcher and discrepancies were discussed and resolved. It was acknowledged that elements such as the reading difficulty of a test might be classified as more than one aspect of CRAS so multiple classifications were allowed. Examples of some judgements are given in Table 1.

To make and quantify the judgements, the data were divided into units. For some studies like Arlett (2002) each questionnaire item could be used as a unit. Each row in Table 1 represents a unit.

Limitations

Inevitably there is some subjectivity involved in the unitisation and the judgements, and other researchers might have come to somewhat different decisions. Nonetheless, the present study is a credible way of investigating the utility of the CRAS framework for comparability studies about VQ/VRQs.

Findings

The results of the study are presented in Table 1 and Table 2. The majority of the data corresponded to an aspect of the CRAS frame of reference.

However, there were some data which did not map to CRAS, but which were classified as a task and/or specification demand(s). For instance, "More general capabilities such as the ability to work in a team" (Coles and Matthews, 1995, p.11), was predominantly affective and interpersonal, whereas CRAS is primarily concerned with the cognitive. Whilst these classifications are assigned to the minority rather than the majority of the data, they are arguably significant in VQ/VRQs. Therefore, using CRAS for comparability studies for VQ/VRQs is likely to mean that some task and/or specification demands, which are significant in VQ/VRQs, are not included in the research.

One of the most striking results is that we did not classify up to 39% of the data from Arlett (2002) as task and/or specification demands. Table 3 provides some data that were classified as not being a task and/or specification demand(s), along with the reason for that decision. Our findings confirm those of Pollitt *et al.* (2007) who found that comparability studies often aim to investigate task and/or specification demands when they are actually investigating something quite different. Indeed it suggests that there is a need to disseminate the technical term and definition of task and/or specification demands to assessment professionals, researchers, assessment setters, specification writers and users of assessments. Otherwise communication can become unclear.

[8] Internal assessment is: *A form of assessment where assessment tasks are set and learners' work assessed by the centre, subject to external moderation or verification where appropriate*. (Ofqual, 2008).

Moderation is: *The process through which internal assessment is monitored to ensure that it meets required standards, and through which adjustments to results are made where required to compensate for any differences in standards that are encountered*. (Ofqual, 2008).

[9] GNVQs or General National Vocation Qualifications aimed to provide study for those intending to stay in full time education but who were not deemed able enough for an A-level programme. The specifications included academic education as well as some vocational learning experiences. The assessments were primarily competence based, evidence gathering and portfolio based rather than external examinations. This information is from Savory *et al*. (2003).

[10] The University of Cambridge Local Examinations Syndicate (UCLES) now has the brand name Cambridge Assessment, which was not in use when SCAA (1995) was written.

Table 1: Examples of data from comparability studies and judgements about how they do or do not map to aspects of CRAS

Data and data source	*Complexity*	*Resources*	*Abstractness*	*Task strategy*	*Response strategy*	*'Other' task and/or specification demand(s) not in CRAS*	*Data 'not' considered to be a task and/or specification demand(s)*
"Evaluation: making judgements based on criteria which have been developed for the purpose. Such as the evaluation of the efficiency of a multi step production process" (Coles and Matthews, 1995:12)	✓	✓	✓	✓	✓		
"How much opportunity is provided for candidates to apply knowledge in their answers to the question paper? A little to a lot" (Arlett, 2002: 12)	✓	✓	✓	✓	✓		
"How specific is the breakdown of marks in the mark schemes? Less specific to more prescribed" (Arlett, 2002: 14)							✓

Note: A tick indicates that the data correspond with an aspect of CRAS.

Table 2: Frequency of data from comparability studies about VQ/VRQs that do or do not map to CRAS

Study	*Total*	*Complexity*		*Resources*		*Abstractness*		*Task strategy*		*Response strategy*		*'Other' task and/or specification demand(s) not in CRAS*		*Data 'not' considered to be a task and/or specification demand(s)*	
Coles and Matthews (1995)	13	11	85%	11	85%	11	85%	11	85%	11	85%	2	15%	0	0%
SCAA (1995)	11	9	82%	9	82%	9	82%	7	64%	7	64%	1	9%	1	9%
Arlett (2002)	23	12	52%	8	35%	7	30%	8	35%	8	35%	0	0%	9	39%
Arlett (2003)	35	18	51%	17	49%	11	31%	6	17%	9	26%	3	9%	11	31%
Guthrie (2003)	26	12	46%	12	46%	12	46%	8	31%	10	38%	2	8%	8	31%
QCA (2006a)	8	8	100%	7	88%	6	75%	6	75%	6	75%	0	0%	0	0%

Note: The first column lists the studies which were included in our investigation. The column labelled 'total' gives the total number of units from each study. The remaining columns refer to the classifications – namely the various aspects of CRAS as well as the categories 'other' and 'not'. Each of these remaining columns has two sub-columns, the left hand sub-column indicates the number of units receiving each classification and the right hand sub-column indicates the number of classified units as a percentage of the total number of units in each study. For each unit more than one classification was allowed, and this is why the percentages in each row do not total 100%.

Table 3: Examples of data and the reason why it was not classified as a task and/or specification demand(s)

Data uni	*The reason the data was not classified as a task and/or specification demand(s)*
"Is the number of marks allocated to each question appropriate?" (Arlett, 2002:13).	Essentially this is an issue of whether the mark scheme was well written and mark allocation was appropriate. The actions a task is intended to require of typical members of the target group of learners are not directly affected by the number of marks allocated to the question.
"Does the mark scheme allow for much compensation/ interpretation?" (Arlett, 2002: 14).	This is about style of mark scheme and whether they allow compensation or whether they are criterion referenced. The actions a task is intended to require of typical members of the target group of learners are not directly affected by whether the mark scheme allows compensation or whether it is criterion referenced.
"How helpful are the mark schemes to: Examiners, in ensuring consistency in marking?" (Guthrie, 2003: 12).	This is about the utility of the mark scheme for examiners. The actions a task is intended to require of typical members of the target group of learners are not directly affected by whether the mark scheme is helpful in ensuring consistency of marking or not.
Whether: "the stated objectives of each scheme were met by the materials considered" (SCAA, 1995: 4).	This is about validity. There are various elements to validity and in this case the issue is the correspondence between what is supposed to be and what actually was measured. The actions the specification is intended to require of typical members of the target group of learners are not directly affected by the correspondence between what is supposed to be and what actually was measured.

References

Arlett, S. (2002). *A comparability study in VCE Health and Social Care units 1, 2 and 5. A review of the examination requirements and a report on the cross moderation exercise. A study based on the summer 2002 examinations*. Organised by the Assessment and Qualification Alliance on behalf of the Joint Council for General Qualifications.

Arlett, S. (2003). *A comparability study in VCE Health and Social Care units 3, 4 and 6. A review of the examination requirements and a report on the cross moderation exercise. A study based on the summer 2002 examinations*. Organised by the Assessment and Qualification Alliance on behalf of the Joint Council for General Qualifications.

Barry, K. (1997). An analysis of the relative demands of advanced GNVQ science and A-level Chemistry. *Journal of Further and Higher Education*, **21**, 1, 45–53.

Bloom, B.S. (Ed.) (1956). *Taxonomy of Educational Objectives – Book 1 – Cognitive Domain*. Michigan: Longman.

Coles, M. & Matthews, A. (1995). *Fitness for purpose: a means of comparing qualifications. A report to Sir Ron Dearing to be considered as part of his review of 16–19 education*.

Coles, M. & Matthews, A. (1998). *Comparing qualifications – Fitness for purpose. Methodology paper*. London: Qualifications and Curriculum Authority.

Crisp, V. & Novaković, N. (2009). Are all assessments equal? The comparability of demands of college-based assessments in a vocationally-related qualification. *Research in Post Compulsory Education*, **14**, 1, 1–18.

Edwards, E. & Adams, R. (2003). *A comparability study in GCE Advanced Level Geography including the Scottish Advanced Higher Grade Examinations. A review of the examination requirements and a report on the cross moderation exercise*. A study based on the Summer 2002 Examination and organised by WJEC on behalf of the Joint Council for General Qualifications.

Gagne, R.M. (1985). *The conditions of learning and theory of instruction* (4th Ed.). New York: Holt, Rinehart and Winston.

Guthrie, K. (2003). *A comparability study in GCE business studies units 4, 5, and 6 VCE business units 4, 5, and 6. A review of the examination requirements and a report on the cross moderation exercise*. A study based on the summer 2002 examinations. Organised by EdExcel on behalf of the Joint Council for General Qualifications.

Hughes, S., Pollitt, A. & Ahmed, A. (1998). *The development of a tool for gauging the demands of GCSE and A-level examination questions*. Paper presented at the British Educational Research Association Annual Conference, The Queen's University of Belfast.

Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design*. New York: Methuen.

Learning and Skills Council (2009). Jargon Buster http://www.lsc.gov.uk/Jargonbuster/Vocational+certificate+of+education.htm [Accessed September 2009]

Mitchel, L. & Bartram, D. (1994). The place of knowledge and understanding in the development of National Vocational and Scottish Vocational Qualifications. In: *Competence & assessment briefing series no. 10*.

OCR (2009). http://www.ocr.org.uk/qualifications/type/nvq/index.html [Accessed January 2010]

OCR (2009). http://www.ocr.org.uk/qualifications/type/vrq/index.html [Accessed January 2010]

Ofqual (2008). Glossary http://www.ofqual.gov.uk/501.aspx#I [Accessed January 2010]

Pollitt, A., Ahmed, A. & Crisp, V. (2007). The demand of examination syllabuses and question papers, 166–206. In: P. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.) *Techniques for monitoring the comparability of examination standards*. London: QCA.

QCA (2006a). *Comparability study of assessment practice: Personal license holder qualifications*, QCA/06/2709 http://www.ofqual.gov.uk/files/personal_licence_holder_quals_comparability_study.pdf

QCA (2006b). *Comparability study of assessment practice Door supervision qualifications* QCA/06/2710 [on line] Available at: http://www.ofqual.gov.uk/files/door_supervision_quals_comparability_report.pdf [Accessed September 2009].

QCDA (undated). Glossary http://testsandexams.qcda.gov.uk/15862.aspx#S [Accessed January 2010].

Savory, C., Hodgson, A. & Spours, K. (2003). *The Advanced Vocational Certificate of Education (AVCE): A general or vocational qualification? Broadening the Advanced Level Curriculum*. IoE/Nuffield Series Number 7, School of Lifelong Education and International Development, Institute of Education, University of London [on line] http://www.ioe.ac.uk/schools/leid/nuff/rep7.pdf [Accessed September 2009]

SCAA (1995). *Report of a comparability exercise into GCE and GNVQ business*. London: School Curriculum and Assessment Authority.

ASSURING QUALITY IN ASSESSMENT

Developing and piloting a framework for the validation of A levels

Stuart Shaw CIE Research **and Victoria Crisp** Research Division

Introduction

This article reports briefly on a current strand of research which aims to develop a methodology for validating general academic qualifications such as A levels. Validity is a key principle of assessment, a central aspect of which relates to whether the interpretations and uses of test scores are appropriate and meaningful (Kane, 2006). For this to be the case, various criteria must be achieved, such as good representation of intended constructs, and avoidance of construct-irrelevant variance. Additionally, some conceptualisations of validity include consideration of the consequences that may result from the assessment, such as affects on classroom practice. The kinds of evidence needed may vary depending on the intended uses of assessment outcomes. For example, if assessment results are designed to be used to inform decisions about future study or employment, it is important to ascertain that the qualification acts as suitable preparation for this study or employment, and to some extent predicts likely success.

Validity has long been considered a crucial criterion for an assessment and there now exists a wealth of theoretical work attesting to its importance. However, practical examples of how to validate an assessment are less common largely because "validation work is unglamorous and needs to be painstaking" (Wood, 1991, p.151–2). To *validate* an assessment, evidence to support the claims made about the assessment must be provided. Providing appropriate evidence for validity is not a simple undertaking and requires multiple sources of evidence collected through a range of methods (Bachman, 1990). This allows different facets important to validity to be addressed and can thus support claims for the validity of scores on an assessment.

The current work focuses on Kane's (2006) definition which states that validity is about the extent to which the inferences made on the basis of the assessment outcomes are appropriate. Given that a key inference is usually that the scores reflect ability or attainment in relation to a particular predefined set of knowledge, understanding and skills, evaluating validity will include considering whether the assessment is measuring what it was intended to measure. Cambridge Assessment sees a vital aspect of validity as "the extent to which the inferences which are made on the basis of the outcomes of the assessment are meaningful, useful and appropriate" (2009, p.8) and argues that the concern for validation "begins with consideration of the extent to which the assessment is assessing what it is intended to assess and flows out to the uses to which the information from the assessment is being put" (2009, p.8).

A debated issue in validity theory is whether the social and personal consequences of assessments should be included within the conceptualisation of validity. This includes issues such as backwash onto classroom practices, and the consequences for individual students of assessment outcomes being used in particular ways. A number of key theorists, including Kane (2006) and Messick (1989) include consideration of consequences within the notion of validity. However, this is somewhat problematic in how it relates to the definition of validity, since not all types of consequences can be considered to relate to the appropriateness of interpretations and uses of test scores. For example, consequences in terms of classroom practices which prepare students for examinations do not relate directly to uses or interpretations of scores. Nonetheless, the consequences are agreed to be important, and arguably fall within a broader notion of the validity of assessment systems and associated curricula. An assessment agency cannot be held responsible for all possible uses of the outcomes of its assessments, but it can take responsibility for being very clear regarding legitimate uses and provide appropriate guidance.

The current line of research aimed to design a set of methods for validating UK qualifications such as A levels and their international counterparts. It is intended that these can later be used on a routine basis or as part of an ongoing validation programme. As the methods need to be underpinned by theoretical understandings of validity, relevant literature was reviewed to develop a standpoint from which to work. There are significant challenges in doing this, not least because of issues around the conceptualisation of validity to be taken and the boundaries of what should be considered in a validation study.

A number of frameworks for validation have previously been proposed (e.g. Cronbach, 1988; Frederiksen and Collins, 1989; Linn, Baker and Dunbar, 1991; Messick, 1989; 1995; Crooks, Kane and Cohen, 1996; Mislevy, Steinberg and Almond, 2002; Shaw and Weir, 2007). However, these tend to involve substantial technical language, to sometimes be specific to particular assessment contexts, and often fail to suggest a set of methods to be used.

Our aim was to develop a comprehensive framework for validation that includes aspects from key theoretical models, but is more accessible and provides an associated set of methods (though the exact methods to be used may vary depending on the nature of the assessment to be validated).

Framework development

This research began by drawing on existing models for validation in various contexts to develop a new framework by which to structure validation exercises for general qualifications. This framework takes the form of a list of validity questions, each of which is to be answered by the collection of relevant evidence. The validity questions are structured within three areas as shown in Figure 1. The findings of validation exercises based on the framework would present '*Evidence for validity*' and any potential '*Threats to validity*'. Any identified threats to validity might provide advice for test development in future sessions, or might suggest recommendations for changes to an aspect of the qualification, its administration and procedures or associated documentation. For a full description of the development of the framework please see Shaw, Crisp and Johnson (2009).

Figure 1: Validation framework questions

1. Assessment purpose(s) and underlying constructs

1.1) What is (or are) the main declared purpose(s) of the assessment and are they clearly communicated?

1.2) What are the constructs that we intend to assess and are the tasks appropriately designed to elicit these constructs?

1.3) Do the tasks elicit performances that reflect the intended constructs?

2. Adequate sampling of domain, reliability and generalisability

2.1) Do the tasks adequately sample the constructs that are important to the domain?

2.2) Are the scores dependable measures of the intended constructs?

3. Impact and inferences

3.1) Is guidance in place so that teachers know how to prepare students for the assessments such that negative effects on classroom practice are avoided?

3.2) Is guidance in place so that teachers and others know what scores/grades mean and how the outcomes should be used?

3.3) Does the assessment achieve the main declared purpose(s)?

The intention is that by collecting evidence relating to each of the components of validity represented by the questions in the framework, an awarding body can provide justification for the validity of its assessments. The aim is to move towards a set of methods that can be operationalised periodically for all of an awarding body's qualifications. Thus, an initial set of methods was devised drawing, where possible, on previous relevant research methods. By facilitating the collection of evidence relating to each question in the framework, the methods give a view of the extent to which the interpretations and uses of an assessment can be considered valid. Multiple sources of evidence are required in order to provide proof that certain inferences are justified.

Piloting with A level Geography

The provisional set of methods was piloted on the assessments involved in an A level geography syllabus which is available internationally. This A level is assessed through three written exam papers.

The piloting used a broad set of methods to explore the different validity questions in the framework. For practical reasons, it would not be possible to use all of these methods operationally for all of an awarding body's qualifications, but this pilot intentionally employed more methods than might normally be practical in order to identify which are most valuable in providing validity evidence.

The set of methods used involved:

- a series of tasks conducted by geography experts (four senior examiners and two external experts) such as identifying assessment constructs, rating the coverage of Assessment Objective subcomponents, and rating the demands of tasks;
- document reviews, for example, in relation to guidance on teaching practice;
- statistical analyses of item level data, including Rasch analysis;
- a multiple re-marking study, involving five markers for each paper, to explore marking reliability;
- questionnaires to teachers and to higher education institutions;
- interviews with students after they had answered example exam questions.

The various methods and analyses allowed consideration of the evidence in relation to each of the questions in the framework for A level Geography. For each, evidence for validity and any possible threats to validity could be identified. For example, a sample of scripts was obtained and the scores were analysed using various statistical methods including Item Response Theory. This provides some evidence relating to question 1.3 in the framework (see Figure 1) about whether the assessment measures the intended constructs. This offered the following insights:

- *Evidence for validity* – Few excessively easy, excessively difficult or misfitting questions were identified. Additionally, the difficulty measures for different optional questions were fairly similar, suggesting reasonable comparability.
- *Possible threats to validity* – One question part showed clear (but slight) misfit for a number of reasons.

To give another example, the questionnaire to teachers included questions about the intended meaning and uses of scores and grades and guidance provided by the examination board, thus relating to validity question 3.2 in the framework. The evidence this provided can be summarised as follows:

- *Evidence for validity* – Teachers reportedly knew how to use exam scores/grades to inform their teaching. Most teachers felt that the guidance available helped them advise students on their future education and/or employment.
- *Possible threats to validity* – Some teachers felt that more guidance could be available on the meaning and use of scores/grades.

The available evidence, from all methods and analyses, were later synthesised in order to provide an overall evaluation of the validity argument. Overall, the findings from the piloting with A level Geography suggest substantial support for the validity of the assessments. However, there were a few minor areas of concern which should be addressed to further increase the validity of the qualification's assessments. These issues have been fed back to the examining team and relevant assessment personnel.

Revising the framework

A further, ongoing, phase of this research aims to build on and refine the framework and methods, in order to move towards a validation model that is more manageable on a routine basis or as part of a long term monitoring programme considering different qualifications and subjects.

The experience of the piloting, feedback and discussion with colleagues and further consideration of the literature on validity has led to refinement of the framework. Changes have been made in relation to how it deals with assessment purposes and also in relation to evaluating qualifications as preparation for future study, if they are used for selection purposes.

Applying a revised set of methods to A level Physics

The set of methods used in the pilot with A level Geography has been revised to give a streamlined subset of methods. Methods have been selected on the basis of how useful they were in providing evidence to evaluate validity and based on their practicality. In addition, some revisions have been made to the previously used methods in light of experience, and one additional method has been added to reflect changes to the framework.

The revised set of methods is currently being used with International A level Physics, to provide evidence to support the claim for its validity, and to identify any potential threats to validity for this qualification such that they can be addressed.

Reflections on the work so far

This project so far has made progress in developing a framework for validation that is suitable for traditional written examinations and in showing that this can be applied to assessments through use of a variety of methods and analyses. This research has also highlighted the challenges faced when validating the intended interpretation of test scores and their relevance to the proposed uses of those scores. These challenges include issues relating to:

- the view of validity adopted and its boundaries;
- the scope and sufficiency of evidence;
- balancing operational manageability and comprehensiveness of evidence;
- a possible need for additional frameworks and sets of methods for different types of qualifications and assessments.

It is hoped that the continuation of this research will help resolve some of these challenges and provide a way forward. Eventually, it is proposed that validation evidence will be collected and presented in an operationally-orientated portfolio for any one particular qualification. This will show more clearly how an appropriate methodology can be used as part of regular monitoring of assessment validity.

References

Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Cambridge Assessment. (2009). *The Cambridge Approach. Principles for designing, administering and evaluating assessment*. Available online at: http://www.cambridgeassessment.org.uk/ca/digitalAssets/181348_cambridge_approach.pdf

Cronbach, L.J. (1988). Five perspectives on validity argument. In: H. Wainer and H. Braun (Eds.), *Test Validity*. 3–17. Hillsdale, NJ: Lawrence Erlbaum.

Crooks, T.J., Kane, M.T. & Cohen, A.S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy and Practice*, **3**, 3, 265–286.

Frederiksen, J.R. & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, **18**, 9, 27–32.

Kane, M.T. (2006). Validation. In: R.L. Brennan (Ed.), *Education Measurement* (4th ed.). Westport: Praeger.

Linn, R.L., Baker, E.L. & Dunbar, S.B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher*, **20**, 8, 15–21.

Messick, S. (1989). Validity. In: R. Linn (Ed.) *Educational Measurement*. 13–103. New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, **50**, 741–749.

Mislevy, R.J., Steinberg, L.S. & Almond, R.G. (2002). Design and analysis in task-based language assessment. *Language Testing*. Special Issue: Interpretation, intended uses, and designs in task-based language, **19**, 4, 477–496.

Shaw, S.D., Crisp, V. & Johnson, N. (2009). *A proposed framework for evidencing assessment validity in large-scale, high-stakes international examinations*. A paper presented at the Association for Educational Assessment in Europe, 10th Annual Conference, Malta, November 2009.

Shaw, S.D. & Weir, C.J. (2007). *Examining Writing: Research and Practice in assessing second language writing*. Cambridge: Cambridge University Press.

Wood, R. (1991). *Assessment and Testing: A survey of research*. Cambridge: Cambridge University Press.

EXAMINATIONS RESEARCH

Statistical Reports

The Statistics Team Research Division

The ongoing 'Statistics Reports Series' provides statistical summaries of various aspects of the English examination system such as trends in pupil attainment, qualifications choice, combinations of subjects and subject provision at school. These reports, produced using national-level examination data, are available in .pdf format on the Cambridge Assessment website: http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Statistical_Reports

The following reports have been published since Issue 9 (January 2010) of *Research Matters*:

- Statistics Report Series No.14: A-level candidates attaining 3 or more 'A' grades in England, 2006–2009
- Statistics Report Series No.15: Provision of science subjects at GCSE, 2009
- Statistics Report Series No.16: A-level uptake and results by gender, 2002–2008
- Statistics Report Series No.17: GCSE uptake and results by gender, 2002–2008
- Statistics Report Series No.18: A-level uptake and results by school type, 2002–2008
- Statistics Report Series No.19: GCSE uptake and results by school type, 2002–2008

Research News

Cambridge Assessment Network

5th Cambridge Assessment Conference: Challenges of Assessment Reform

21 October 2010, Robinson College, Cambridge

Assessment is under reform all over the world. Some countries are embracing national testing, while others are abandoning it. Many countries are struggling to understand how best to integrate Assessment for Learning within everyday classroom practice.

The potential gains from reform may be high but the processes of change are complex and the consequences of getting it wrong can be severe. In the past decade alone, England has witnessed numerous crises of assessment reform; from the introduction of Curriculum 2000 A levels and the grading furore of 2002, to the appointment of a new contractor for marking national curriculum tests and the marking furore of 2008. Current reforms include the introduction of stretch and challenge at A level, diploma qualifications, revised arrangements for national testing, functional skills testing, controlled assessments, project qualifications, and more.

The 5th Cambridge Assessment Conference will address the challenges of assessment reform. What ensures its success? What undermines it? What lessons can we learn from reforms past and present?

Cambridge Assessment is pleased to announce that Professor Paul Black, King's College London, will be opening the conference with a keynote presentation on the effective integration of pedagogy, learning and assessment, as the foundation for successful assessment reform. The conference will include a panel discussion chaired by Mike Baker, former BBC Education Editor, and featuring Dr Mary Bousted from the Association of Teachers and Lecturers, Kathleen Tattersall from Ofqual, and others. Our speakers include Professor Jo-Anne Baird from the University of Bristol, Professor Frank Ventura from the University of Malta, Professor Peter Tymms from the University of Durham, as well as additional speakers yet to be confirmed.

Further details of the conference, including details of the programme and how to book your place, will be found on www.assessnet.org.uk/conference2010 as they become available.

To join the mailing list for conference updates, please email us at thenetwork@cambridgeassessment.org.uk, with 'Conference updates' in the subject line.

Seminars

Critical Thinking – skills for life

On 11 February over 60 teachers, industry representatives and leading academics came together at a Cambridge Assessment seminar to discuss the issue of Critical Thinking and whether it should be treated as a specialist, stand-alone subject or 'embedded'.

The seminar took place at the British Academy and was held to highlight how an explicit focus on Critical Thinking can enhance the attainment of pupils of all backgrounds and abilities, following recent research undertaken by Cambridge Assessment Senior Research Officer, Beth Black. Importantly, the research showed that pupils who study Critical Thinking as a discrete subject at AS level tend to do better in their other A level subjects, whether they are taking sciences, languages or humanities.

Although the debate was wide ranging, there was widespread agreement about the benefits of thinking skills in education and employment and a consensus that, whether delivered separately or embedded, it is important that the teaching of Critical Thinking be explicit.

Further details can be found at: http://www.cambridgeassessment.org.uk/ca/Spotlight/Detail?tag=Critical

For a personal response to the seminar see Joe Chislett's article on p.9.

Publications

The following articles have been published since Issue 9 of *Research Matters*:

Crisp, V. (2010). Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education*, **26**, 1, 1–21.

Crisp, V. (2010). Judging the grade: exploring the judgement processes involved in examination grading decisions. *Evaluation and Research in Education*, **23**, 1, 19–35.

Johnson, M. and Crisp, V. (2010). A case of positive washback: an exploration of the effect of pre-release examination materials on classroom practice in the UK. *Research in Education*, **82**, 47–50.

Johnson, M. and 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., Nádas, R. and Green, S. (2010). Marking essays on screen and on paper. *Education Journal*, **121**, 39–41.

Nádas, R. and Suto, I. (2010). Speed isn't everything: a study of examination marking. *Educational Studies*, **36**, 1, 115–118.

Suto, I. and Shiell, H. (2009). What influences moderation and standards maintenance in school-based summative assessment? *Education Journal*, **119**, 41–43.

For all the latest research news please visit http://www.cambridgeassessment.org.uk/ca/Our_Services/Research

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

Tel: 01223 553854
Fax: 01223 552700
Email: ResearchProgrammes@cambridgeassessment.org.uk
http://www.cambridgeassessment.org.uk

ISSN: 1755–6031