Issue 2 June 2006

Research Matters

CAMBRIDGE ASSESSMENT

University of Cambridge
Local Examinations Syndicate

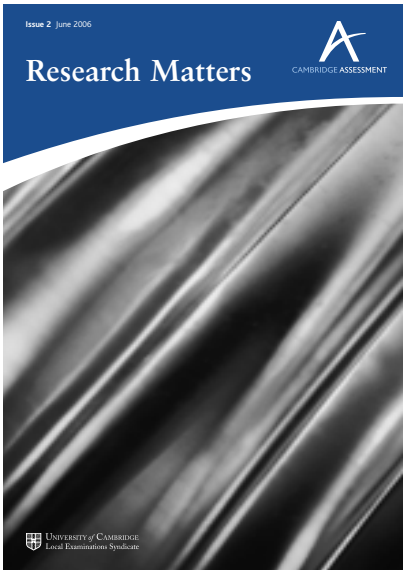
CAMBRIDGE ASSESSMENT

Citation

Articles in this publication should be cited as: Crisp, V. and Johnson, M. (2006). Examiners' annotations: Practice and purpose. *Research Matters: A Cambridge Assessment Publication,* 2, 11-13.

Research Matters : 2

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 copies of articles are available on our website
www.cambridgeassessment.org.uk/research

Foreword

Welcome to the second issue of *Research Matters*, a biannual publication from Cambridge Assessment. The aim of this publication is to share assessment research in a range of fields with colleagues within Cambridge Assessment and in the wider assessment community and to comment on prominent research issues. Contributions are mostly from Assessment Research and Development staff working across many areas of assessment and examinations. However, in this issue we also include two articles from the *Financial Times* which take an interesting view of A-level examinations. Since the publication of Issue 1 we have published a special issue which reported on *Variations in aspects of writing in 16+ English examinations between 1980 and 2004*. I hope that you will find *Research Matters* interesting and informative.

I am pleased to announce that Tim Oates has joined us as Group Director of the growing Assessment Research and Development Division. Tim joined us from the Qualifications and Curriculum Agency, where he was Head of Research and Statistics for most of the last decade. Tim is ideally qualified to make sure that research is at the heart of Cambridge Assessment and I am delighted to welcome him on board.

Simon Lebus *Group Chief Executive*

Editorial

In this issue we report on a wide range of research topics from vocational assessment to new technologies. The first four articles describe research presented at conferences. In the opening article Irenka Suto and Sylvia Green discuss international perspectives on vocational education. Much of this project was presented in a discussion group at the annual conference of The Association for Educational Assessment in Europe (AEA-Europe) in November 2005. This is followed by an article by Irenka Suto and Jackie Greatorex on 'A cognitive psychological exploration of the GCSE marking process'. Aspects of this research were presented at two conferences in 2005: those of the British Educational Research Association (BERA) and AEA-Europe. 'Examiners' annotations: Practice and purpose' by Victoria Crisp and Martin Johnson looks at current annotation practices in examination marking and was presented at the BERA conference in September 2005. Martin Johnson and Jackie Greatorex then discuss issues of validity in judging learners' work on screen, which was also presented at the BERA conference.

These articles are followed by an evaluation of the Cambridge Assessment/Oxford University automatic marking system by Nicholas Raikes which follows from his article in the first issue of *Research Matters*. John Bell and Joanne Emery then take a new look at the debate surrounding the decline in the number of students taking A-level mathematics and discover that the situation is more optimistic than has been reported.

The two articles from *The Financial Times*, reprinted with their permission, describe what happened when four *Financial Times'* columnists agreed to sit some of the 2005 papers.

In his article, 'The Cambridge Assessment Network', Andrew Watts outlines the way in which the Research Division and Cambridge Assessment Network work together to enhance professional development in the assessment community.

And finally, 'Research News' includes conferencing information and details of recent publications.

Sylvia Green *Director of Research*

International perspectives on vocational education: What can we learn from each other?

Dr Irenka Suto and Sylvia Green Research Division

Introduction

The broad aim of Vocational Education and Training (VET) is to provide students with the technical skills and knowledge needed to enter the workforce. It exists for an extensive range of subject areas and may be delivered through many different kinds of training institutions and enterprises. Over the past four years, this huge area of education and assessment has received close scrutiny from the Government and others, as part of a broader review of education and skills among 14 to 19 year olds (Tomlinson, 2004; Department for Education and Skills, 2005; Fuller and Unwin, 2005; Hodgson and Spours, 2005). Given that this process has resulted in proposals for considerable reform of VET for this age group, we deemed it important to know more about the international context within which they are set. Who does VET affect globally, and what might we learn from the experiences of other countries? The aims of this project, therefore, were to identify and examine two main types of data: (i) on the extent of participation in VET and its associated assessment worldwide; and (ii) relating to key differences in the VET systems of different countries.

There were three stages to the project:

1. A review of the quantitative data available.
2. A review of other key literature.
3. A discussion group at an international conference.

In this report, we summarise some of the main findings from each stage.

1. Review of the quantitative data available

Questions

Pass rates, enrolment figures, and methods of assessment for *general* qualifications, such as GCSEs and A-levels, are frequent topics of debate in the media. Large quantities of data are collated each year and are made available for analysis. Vocational qualifications, however, have received less attention of this kind, both from the public and from professionals. We began this review by identifying the following key questions for consideration:

1. What proportions of upper secondary school students participate in vocational streams of education (compared with general streams) in the UK and in other countries?
2. By what means are these students assessed?

Definitions

It was important from the start to set out the definitions and boundaries to be used in the review when searching for data. Since in most countries, enrolling in VET is not possible before the end of compulsory education (Cedefop, 2003), we chose to focus on participation at level 3 of UNESCO's International Standard Classification of Education (ISCED) 1997, which is termed *Upper Secondary Education*. The principal characteristics of ISCED level 3 are as follows:

> *'This level of education typically begins at the end of full-time compulsory education for those countries that have a system of compulsory education. More specialization may be observed at this level than at ISCED level 2 and often teachers need to be more qualified or specialized than for ISCED level 2. The entrance age to this level is typically 15 or 16 years.*
>
> *The educational programmes included at this level typically require the completion of some 9 years of full-time education (since the beginning of level 1) for admission or a combination of education and vocational or technical experience and with as minimum entrance requirements the completion of level 2 or demonstrable ability to handle programmes at this level'* (ISCED, 1997, paragraphs 62–63).

Although the age range of children included in this UNESCO level begins slightly higher than that used in the UK (for example, in the Tomlinson review), we used it in order to ensure that international comparisons would be possible. Educational programmes at ISCED level 3 can be sub-classified along three dimensions:

(i) whether the orientation of the programme is general or vocational;

(ii) the destinations for which the programme has been designed to prepare students; and

(iii) cumulative theoretical duration in full time equivalent since the beginning of ISCED level 3.

The typical duration of programmes ranges from two to five years.

A further boundary to the review was the recency of the data to be considered. We restricted our search to data collected from 1997 onwards, the date of UNESCO's most recent ISCED definitions. A final boundary was our decision to focus on the twenty-five countries in the European Union. However, although exploring VET in all countries was judged to be beyond the scope of this project, we also obtained some data from some other key 'developed' countries where it was readily available (Australia, Canada, China, Japan, New Zealand, and the USA).

Data sources

The following sources of information were investigated:

1. International organisations

Several international organisations provide core data on their websites, as well as downloadable reports of their own research and experiences of data collection. We explored the websites of UNESCO, OECD, Eurostat, and Cedefop (the European Centre for the Development of Vocational Education and Training, which is the EU's reference centre for VET).

Several relevant reports were obtained in this way, sometimes through direct contact with representatives of these organisations.

2. Government departments

Departmental and related websites were searched for data on VET. In one case (the Republic of Ireland) direct contact was made with civil servants to confirm unusual information obtained from international organisations.

3. Research literature and related publications

Published and 'grey' research literature were searched using the on-line facilities of the Cedefop library (based in Thessaloniki), and also with the assistance of its librarians. This is Europe's largest specialist VET library, housing multilingual collections, research reports, EU legislation, and comparative studies. To supplement this search, a number of internationally respected peer-reviewed journals were hand-searched for data and potential leads to other data sources. The Educational Resources Information Center (ERIC) on-line database was also searched.

Findings

The most striking finding of the review was the paucity of quantitative data on the topics of interest. The reasons for this are considered subsequently, in the second stage of the project (the review of other key literature). Nevertheless, it was still possible to obtain data in several interesting areas.

Enrolment of ISCED 3 students in general and vocational streams

As shown in Figure 1, the proportion of upper secondary students enrolled in a vocational rather than a general stream of education varies considerably from country to country, within the EU and also worldwide. Among the countries with the lowest proportions of VET enrolments are Canada, Hungary and Cyprus, where fewer than 15% of upper secondary students are enrolled. At the other end of the spectrum are the Czech Republic, Slovakia, Austria, Belgium, Slovenia and the UK, where over two thirds of enrolments are in vocational streams. In Ireland and the USA, all students are enrolled in general streams of education because there are no formal upper secondary vocational streams at ISCED level 3. (According to statisticians at the Department of Education and Science in Ireland (O'Rourke and Dunne, personal communication), the lowest coding for Irish vocational students, for example apprenticeships, is at ISCED level 4.)

There are some striking differences in the enrolment figures of some countries that neighbour each other and might therefore have been presumed to have similar educational systems. For example, Australia's VET enrolment rate of 63% contrasts starkly with New Zealand's rate of just 24%. Poland has a VET enrolment rate of 54%, which is somewhat lower than those of neighbouring Slovakia (75%) and the Czech Republic (79%).

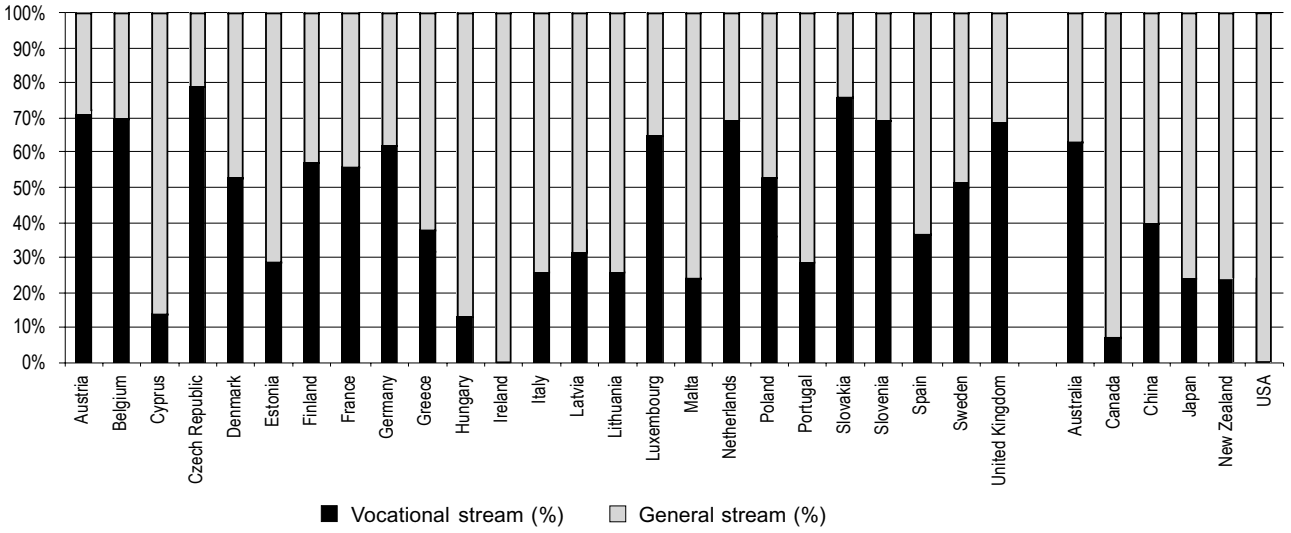
As Figure 2 reveals, in all EU countries except the UK, there are more males than females enrolled in the vocational stream of upper secondary education; on average, the difference in VET enrolment rates is 10.5%. However, these differences vary considerably in their magnitude from country to country. In Italy, the male and female enrolment rates differ by just 3.5%, in Belgium the difference is 3.7%, and in the Netherlands it is 4.0%. In Estonia on the other hand, the difference in enrolment rates is as great as 20.9%, and there are also large differences in Poland (19.2%), Malta (18.9%) and Cyprus (17.6%).

Assessment and other completion requirements for VET

In most of the countries considered, the variety of VET available is quite substantial. There are many different types of programme, which are organised by a range of training institutions and organisations, and which combine tuition with work-based experience to varying degrees. With such variety comes a range of systems of summative assessment and other requirements for programme completion. Although data are sparse, the OECD has categorised the educational programmes of some countries according to whether they require students to attend a specified number of course hours, and whether students are examined on their achievements (for full details, see www.oecd.org/edu/eag2005).

According to this (partial) data, the completion only of a specified number of course hours is a rare requirement, occurring in Korea alone

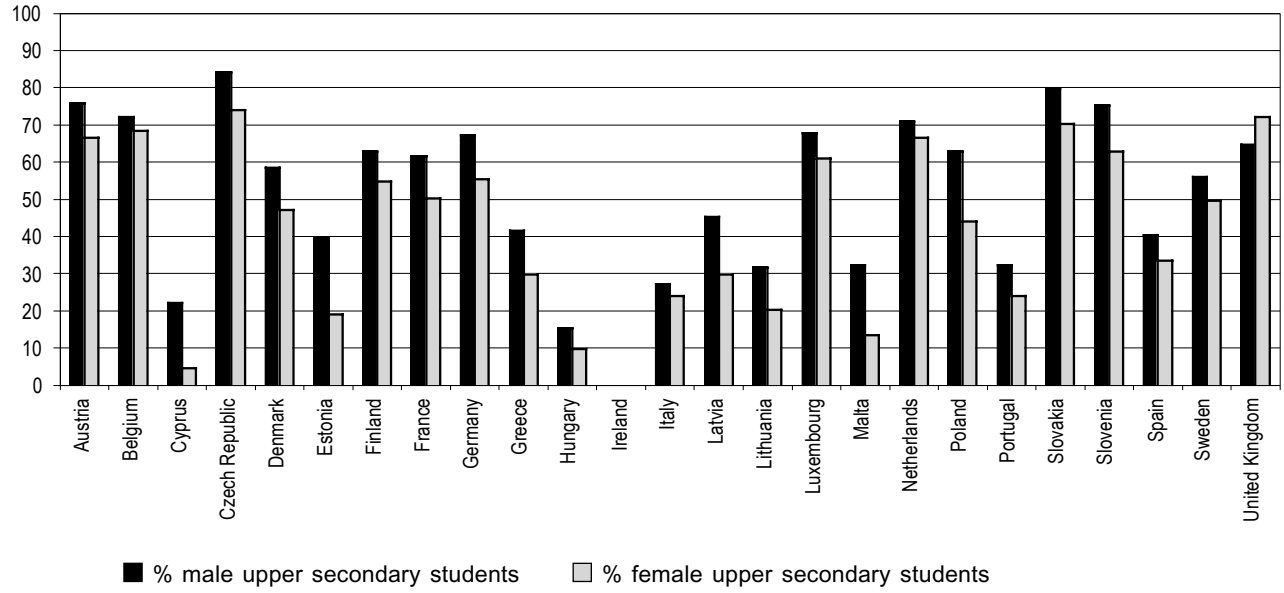
Figure 1: Enrolments of upper secondary (ISCED 3) students in general and vocational streams in the EU and in six other major countries



Note: According to statisticians at the Department of Education and Science in Ireland, there are no Irish vocational students at ISCED level 3. The lowest coding for Irish vocational students, for example apprentices, is at ISCED level 4.

Data sources: Eurostat, UNESCO Institute for Statistics. All data relate to 2003.

Figure 2: Percentages of male and female upper secondary (ISCED 3) students enrolled in vocational streams in the EU



Note: According to statisticians at the Department of Education and Science in Ireland, there are no Irish vocational students at ISCED level 3. The lowest coding for Irish vocational students, for example apprentices, is at ISCED level 4.

Data sources: Eurostat, UNESCO Institute for Statistics. All data relate to 2003.

for programmes most likely to have a vocational orientation. Several countries, including Austria, the Czech Republic, Hungary, Japan, and Luxembourg, have ISCED 3 programmes of all types for which students are required both to attend a specified number of course hours and to pass some form of examination(s). In Iceland, the completion of a specified number of course hours is not a requirement for any ISCED 3 students; the implication is that passing examinations is the sole requirement. Italy, on the other hand, relies solely on examinations for some ISCED 3 programmes but not others.

The OECD has also collated data on whether ISCED 3 students are given final examinations and/or a series of examinations during their programmes. For both ISCED 3B and ISCED 3C programmes (i.e. those programmes most likely to have a vocational orientation), over two thirds of countries (for which there are data) have a series of examinations during at least some programmes, and approximately one third of countries have some programmes that do not make use of such examinations. Roughly two thirds of the countries have some ISCED 3B and ISCED 3C programmes entailing final examinations, and well over one third of countries have some programmes that do not. For ISCED 3A (i.e. largely theoretically based) programmes, the proportions of countries making use of final examinations and series of examinations during their programmes are broadly similar.

For programmes with a vocational orientation, no internationally comparable data could be found on the characteristics of the examinations that take place. For example, to date, there have been no major international quantitative comparisons of the usage of internal and external examiners and the training that they receive, or of the extent to which examinations are based upon practical performance. Similarly, statistics enabling the reliability and validity of assessments in vocationally orientated programmes to be compared internationally were unobtainable. This may well be because the diversity of programmes, even within each country, makes such tasks extremely difficult.

2. Review of other key literature

The first stage of this project exposed a lack of internationally comparable quantitative data relating to VET. However, it also revealed some interesting and relevant publications in some related areas, two of which are worth summarising:

(i) a major systematic literature review entitled *What determines the impact of vocational qualifications?* conducted for the DfES (Unwin, Fuller, Turbin and Young, 2004); and

(ii) research on competency-based assessment in Australian VET.

What determines the impact of vocational qualifications?

During 2003, Unwin and her colleagues conducted an extensive literature review, identifying and scrutinising 'primary' (empirical research) studies, 'secondary' studies (analyses of major data sets), conceptual studies, academic critiques, and policy documents. The review had multiple aims, but its authors were forced to conclude:

> *'The research-based literature on VQs is thin, reflecting the invisibility of vocational education and the work-based pathways more generally. Where they do exist, studies of vocational education and vocational learning often do not include any focus on VQs, and hence, the UK lacks a substantive evidence base on VQs'* (Unwin *et al.*, 2004, p. 4).

This conclusion is in line with a recent internal review of the literature (Johnson, 2005). Nevertheless, the report contains a useful section (4) in which Unwin and her colleagues offer an international perspective on the development of vocational qualifications in England. The authors argue that the huge variety in what is regarded as VET internationally can explain why it can be so difficult to collect quantitative data and make direct international comparisons, or to make judgements and generalisations about the effectiveness of particular systems and models of VET assessment. Consequently, it is proposed that the best way to use different

national VET models may be as 'mirrors' for examining the assumptions on which our own vocational qualifications are based. A qualitative analysis of national differences may provide a way of making explicit the context within which vocational qualifications have developed.

Unwin and her colleagues focussed on the experiences of several groups of countries: those in continental Europe, North America, South East Asia, Australia and New Zealand, and Scotland. They identified six broad contextual features that affect how vocational qualifications have developed in different countries:

1. The role of the state and other stakeholders
2. The use of vocational qualifications
3. The role for private [independent] Awarding Bodies
4. The relationship between vocational and professional qualifications
5. The role of employers and social partners
6. The extent of license to practice.

Four key ways in which the design and function of vocational qualifications differ among countries were also described:

1. The relationship of qualifications to provision
2. The role of outcomes
3. The development of a national qualifications framework
4. The relationship between vocational and general (academic) qualifications.

All of these different ways in which VET varies internationally are discussed in full by Unwin *et al*. (2004).

Competency-based assessment in Australian VET

Another body of international literature was identified as potentially having particular relevance for VET reforms in England. Over the past few years, competency-based assessment has become very popular in Australia (Williams and Bateman, 2003). There (and elsewhere), an interesting debate has arisen over whether it is meaningful and useful to grade competency-based assessments or whether it is best just to class students as 'competent' or 'not competent'. In their review of the literature, Williams and Bateman (2003) have identified the following main arguments for graded assessment:

- We need to provide more comprehensive information on performance for students themselves, but also for higher education institutions and potential employers. This could lead to fairer and more meritocratic selection processes.
- The possibility of obtaining higher grades may have a motivational impact on students and also their trainers.
- There may be positive effects on teaching and learning, for example, through providing more detailed feedback.
- Similarly, grading gives better feedback to employers about their employees' progress.
- Through grading, aptitudes for specialisation can be recognised.
- Grading enables better validity and reliability estimates to be made.
- Grading generates competition, which is essential for business but also in other work places.

Arguments against graded assessment include the following:

- Grading is not compatible or commensurate with competency-based training:

  'As any particular national competency standard defines only one level of performance, it can be argued that only one standard applies to the assessment and reporting of performance. In other words, competency standards do not allow for levels, one is either 'competent' or 'not competent'.' (Thomson, Mathers and Quick, 1996, p. 10).

- A single cut-off point in competency-based assessment may support access and equal opportunity.
- Grading can lead to a sense of failure among struggling learners. Mature learners should not have to repeat their earlier experiences of failure. Single cut-off points might encourage and boost the confidence of candidates with no chance of getting a high grade.
- Grading can stress and pressurise the candidates, preventing them from performing at their best.

Williams & Bateman (2003) also report on their own major empirical study (conducted in 2000) in which 120 Australian VET stakeholders, including students, were interviewed about grading practices. The authors concluded that there exists a huge diversity of grading practices (and views on grading), and that it is not possible to identify general trends, themes or patterns in practice. However, graded assessment in VET should be:

(i) criterion-referenced,

(ii) applied once competence is determined,

(iii) transparent, and

(iv) discretionary.

3. Discussion group at an international conference

To take forward the approach of Unwin *et al*. (2004), that the best way to use different national VET models may be as 'mirrors' for examining the assumptions upon which our own vocational qualifications are based, an international discussion group was convened. It was held at the Annual Conference of the Association for Educational Assessment in Europe and, lasting one and a half hours overall, it comprised four presentations:

1. Irenka Suto set the scene for debate by discussing the variation in definitions of Vocational Education and Training. She then presented a broad international overview of participation figures for VET. Finally, she outlined some of the issues and major areas for discussion that surround the development of vocational qualifications, as had been identified in the research literature.
2. Sylvia Green presented an overview of the VET situation in England. She discussed the main conclusions and recommendations of the Tomlinson report and the recent White Paper (Tomlinson *et al*., 2004; Department for Education and Skills, 2005). These included the Government's proposals for improved vocational routes, with greater flexibility to combine academic and vocational qualifications, specialised diplomas, more participation among employers, and also increased opportunities for work-based training through Apprenticeships, which would be brought into the diploma framework. An explanation of the National Qualifications Framework was then given, together with an outline of some of OCR's vocational qualifications, which illustrate a range of assessment methods.

3. We invited colleagues in The Netherlands to give the third presentation. Elisabeth van Elsen of Esloo Education Group, Peter Kelder of Northgo College, and Jan Wiegers of CITO presented a detailed account of the Dutch educational system, explaining the four 'learning pathways' of pre-vocational secondary education ('VMBO'), and also the various routes through senior secondary vocational education. New routes in Dutch pre-vocational education include programmes that combine different subjects and specialisations. Assessment comprises a mixture of central and school examinations; for vocational subjects, computer-based assessments of the theory supporting practice are used. A key message of this presentation was that a genuine parity of esteem for vocational and general qualifications exists in The Netherlands. VET and general education are nicknamed 'the two royal routes' and are regarded as distinct but equally valuable forms of study.
4. The final presentation was given by John Lewis of the Scottish Qualifications Authority. The key characteristics of Scottish Progression Awards (SPA) and National Progression Awards (NPA) were outlined, but the presentation focussed upon 'Skills for Work' (SfW) programmes. These new courses, which are aimed at pupils in the third and fourth years of secondary education, address both general employability skills as well as specific vocational skills. They involve varied learning environments, entail partnerships between schools, colleges, employers and training providers, and focus on the world of work. The question of what employability skills actually are was raised, and core skills within VET were also discussed.

Each presentation was followed by, or interspersed with, lively discussion, which included conference delegates from a range of countries across Europe. Debate centred around the following themes:

- Different explanations for the lack of quantitative data were considered. Several delegates concurred with the idea that difficulties in reporting data to international bodies may stem from different conceptions and definitions of VET. It was suggested that individuals responsible for reporting their countries' data may not always be aware of these differences, or have the resources to investigate them, which may compound problems. An Irish delegate argued that, despite information to the contrary from the Department of Education and Science in Ireland, Irish students *did* participate in VET at ISCED level 3, but that it was so well integrated into general educational programmes that it could not be quantified meaningfully.
- It became apparent that, in comparison with general education, relatively little research into aspects of VET and its assessment has been conducted. One explanation offered for this was that most researchers have received a general or 'academic' education themselves, rather than a vocational one, and few VET 'experts' are trained in research methods. Another suggested reason was a lack of pressure from stakeholders because, in many countries, they form disparate groups of professionals. Similarly, the diversity of workplaces can create practical difficulties for researchers. It was suggested that in some countries, satisfaction and contentment among stakeholders may have led to a general feeling that research is not needed.
- The issue of parity of esteem between general and vocational education and qualifications arose at several points during the discussion group. Such parity has been achieved in different ways in different countries. In Ireland, for example, a single integrated educational system is highly regarded, whereas in The Netherlands, two distinct but equally prestigious educational routes have been developed through the application of considerable resources to each. The question arose of why, generally speaking, research into VET has been relatively neglected in the UK, and this led on to a discussion of how vocational options can be made more attractive and relevant to potential students. The increasing popularity of particular routes (for example, information technology in the Netherlands and Sweden) was considered.
- Reasons for gender differences in participation rates for vocational and general education were discussed, and various speculations were made. Again, differences in definitions may affect the reporting of data: some professions in which there are notable gender differences, such as nursing and engineering, may be entered through vocational routes in some countries, but through general routes in others.

Summary and conclusions

To summarise, the three stages of this project were: a review of the quantitative data available; a review of other key literature; and a discussion group convened at an international conference. Our aims were to identify and examine two main types of data: (i) on the extent of participation in VET and its associated assessment worldwide; and (ii) relating to key differences in the VET systems of different countries. We were able to obtain some basic information from international organisations government departments, and the research literature. However, we had to conclude that, in general, there is a paucity of internationally comparable quantitative data relating to VET. Reasons for this are likely to include the differences in definitions of VET among countries, and a lack of research in this field, which, in some countries (including the UK), may be due in part to a lack of parity of esteem with general education and assessment. Despite these difficulties, however, we were able to identify some useful information in the research literature: a major systematic literature review conducted for the DfES (Unwin, *et al*., 2004); and research on competency-based assessment in Australian VET. Both provided us with useful insights into VET, and in particular, into the qualitative differences that exist among countries' educational systems, which were considered subsequently in our international discussion group.

Further reading

An additional 109-item bibliography of related literature, compiled as part of the literature review, is available from the authors on request.
A full version of this article is available on *The Association for Educational Assessment in Europe* website: http://www.aea-europe.net/page-180.html

References

Cedefop (2003). *Key figures on vocational education and training*. Luxembourg: Office for Official Publications of the European Communities.

Cedefop (2005). *The European Centre for the development of vocational education and training*. Information available online at http://www.cedefop.eu.int/

Deissinger, T. (2004). Germany's system of vocational education and training: challenges and modernisation issues. *International Journal of Training Research* **2**, 1, 76–99.

Department for Education and Skills (2005). *14–19 Education and Skills*. London: The Stationery Office.

Department of Education and Science (2005). Irish educational information available online at http://www.education.ie/home/

Eurostat (2005). Detailed statistics on the EU and candidate countries available online at http://epp.eurostat.cec.eu.int

Fuller, A. & Unwin, L. (2005). Opinion: Vocational guidance. *Guardian Unlimited*, Tuesday 29th March.

Harris, M. (2003). *Modern apprenticeships: An assessment of the Government's flagship training programme*. Policy paper. London: Institute of Directors.

Hodgson, A. & Spours, K. (2005). Divided we fail. *The Guardian*, Tuesday 1st March.

Johnson, M. (2005). *Vocational review: Issues of validity, reliability and accessibility*. Internal report for Cambridge Assessment.

Organisation for Economic Co-operation and Development (2005). *Education at a glance*. Available online at www.oecd.org/edu/eag2005

Regional Education Indicators Project – PRIE (2002). *Regional report: Educational panorama of the Americas*. Santiago, Chile: UNESCO Regional Office for Education in Latin America and the Caribbean.

Thomson, P., Mathers, R. & Quick, R. (1996). *The grade debate*. Adelaide, Australia: National Centre for Vocational Education Research.

Tomlinson, M. (2004). *14–19 Curriculum and qualifications reform: Final report of the working group on 14–19 reform*. London: DfES Publications.

UNESCO (1997). International Standard Classification of Education (ISCED). Available online at www.unesco.org/education/information/nfsunesco/doc/isced_1997.htm

UNESCO (2003). Institute for Statistics. Data available online at www.uis.unesco.org

Universities and Colleges Admissions Service (2005). *International qualifications 05: for entry into higher education*. Cheltenham: UCAS.

Unwin, L., Fuller, A., Turbin, J. and Young, M. (2004). *What determines the impact of vocational qualifications?* Department for Education and Skills Research, Report No 522.

Williams, M. & Bateman, A. (2003). *Graded assessment in vocational education and training: An analysis of national practice, drivers and areas for policy development*. Leabrook, Australia: National Centre for Vocational Education Research.

A cognitive psychological exploration of the GCSE marking process

Dr Irenka Suto and Dr Jackie Greatorex Research Division

Background

GCSEs play a crucial role in secondary education throughout England and Wales, and the process of marking them, which entails extensive human judgement, is a key determinant in the futures of many sixteen-year-olds. While marking practices in other kinds of examinations have received some serious consideration among researchers (for example, Cumming, 1990; Vaughan, 1992; Milanovic *et al*., 1996; Laming, 1990, 2004; Webster *et al*., 2000; Yorke *et al*., 2000), the judgements made during GCSE examination marking remain surprisingly little explored. The aims of our study, therefore, were to investigate the cognitive strategies used when marking GCSEs and to interpret them within the context of psychological theories of human judgement.

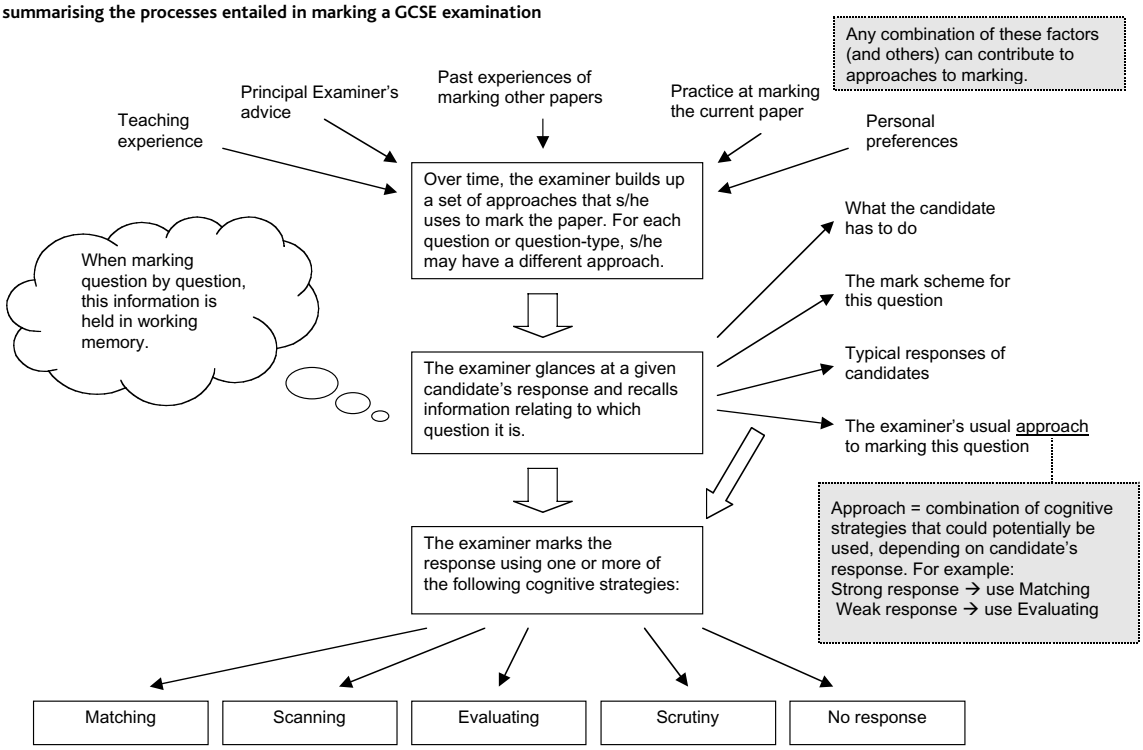
Within the broad field of psychology, there exist multiple models of judgement and decision-making, which have yet to be applied to GCSE examination marking. One potentially useful theoretical approach is that of dual processing. Such models distinguish two qualitatively different but concurrently active systems of cognitive operations: *System 1* thought processes, which are quick and associative, and *System 2* thought processes, which are slow and rule-governed (Kahneman and Frederick, 2002; Stanovich and West, 2002).

The 'intuitive' judgements of System 1 are described as automatic, effortless, skilled actions, comprising opaque thought processes, which occur in parallel and so rapidly that they can be difficult to elucidate (Kahneman and Frederick, 2002). System 2 judgements, in contrast, have been termed 'reflective', and the thought processes they comprise are characterised as slow, serial, controlled, and effortful rule applications, of which the thinker is self-aware (*ibid*. 2002). According to Kahneman and Frederick (2002), as an individual acquires proficiency and skill at a particular activity, complex cognitive operations may migrate from System 2 to System 1. For example, chess masters can develop sufficient expertise to perceive the strength of a chess position instantly, as pattern-matching replaces effortful serial processing.

GCSE examination marking is a diverse activity, encompassing a wide range of subjects with a variety of question styles and mark schemes. It is likely, therefore, that at least some aspects of it will have parallels with some of the activities already scrutinised by judgement researchers in other contexts. There may be question types, or stages of marking, that involve System 1 processing; at times, simple and repetitive matching of a candidate's single-word response with the model answer given in the mark scheme may be all that is required. At other times, examiners might be engaged in System 2 processing; for example, when carefully applying the complex guidelines of a mark scheme to a candidate's uniquely worded essay. As examiners become more familiar with a particular examination paper and mark scheme, or more experienced at marking in general, some sophisticated thought processes may be transferred from System 2 to System 1, while others remain exclusive to System 2.

In the present investigation, we sought to identify and explore some of the many judgements made by GCSE examiners. To do this, we conducted a small-scale empirical study of examiners marking two contrasting subjects, in which we used the 'think aloud' method (Ericsson and Simon, 1993; Leighton, 2004; Van Someren *et al*., 1994) to obtain verbal protocol data for qualitative analysis.

Figure 1: Model summarising the processes entailed in marking a GCSE examination



Methods

Two GCSE examinations (administered by OCR) were considered: an intermediate tier Mathematics paper, which used a 'points-based' marking scheme, and a foundation tier Business Studies paper, which used a 'levels-based' scheme. For both examinations, candidates' scripts comprised individual booklets containing subdivided questions with answer spaces beneath each question part.

For each subject, a group of six experienced examiners (one Principal Examiner and five Assistant Examiners) marked four identical script samples each. The first three of these samples were marked silently (for details, see Suto and Greatorex, *in press*). They were used to familiarise the examiners with the papers and coordinate their marking. Whilst marking the fourth sample (comprising five scripts), the examiners were asked to 'think aloud' concurrently, having been instructed: '*... Say out loud everything that you would normally say to yourself silently whilst you are marking...*' Using a semi-structured interview schedule, the examiners were later questioned about their marking experiences retrospectively.

Results

An extensive qualitative analysis, and interpretation, of the verbal protocol data enabled us to propose a tentative model of marking, which includes five distinct cognitive marking strategies: *matching, scanning, evaluating, scrutinising,* and *no response*. An overview of the model is presented in Figure 1, and the five strategies are presented in detail in Figures 2 to 6. (There is a key to these figures on page 9.) These strategies were broadly validated not only in the retrospective interviews with the examiners who participated in the study, but also by other senior mathematics and business studies examiners.

Figure 2: The 'Matching' strategy

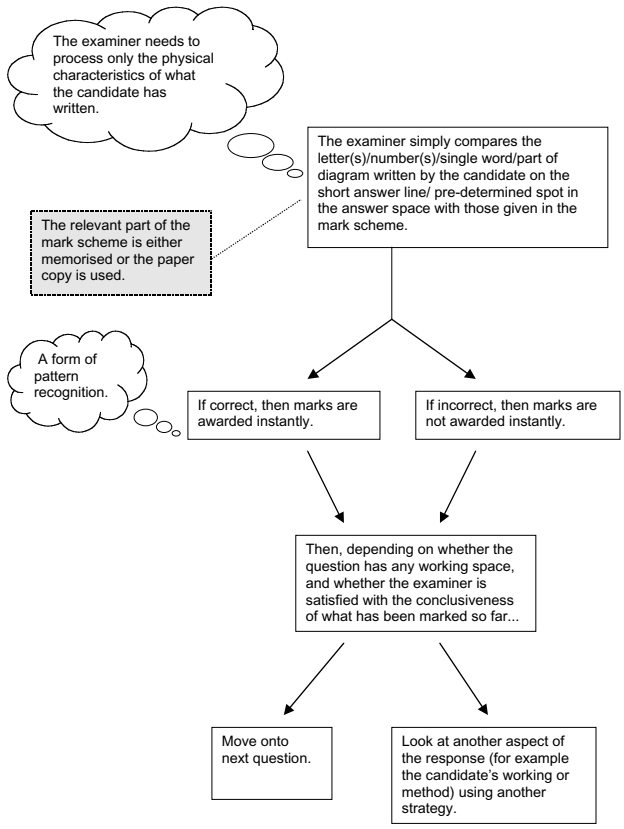


Figure 3: The 'Scanning' strategy

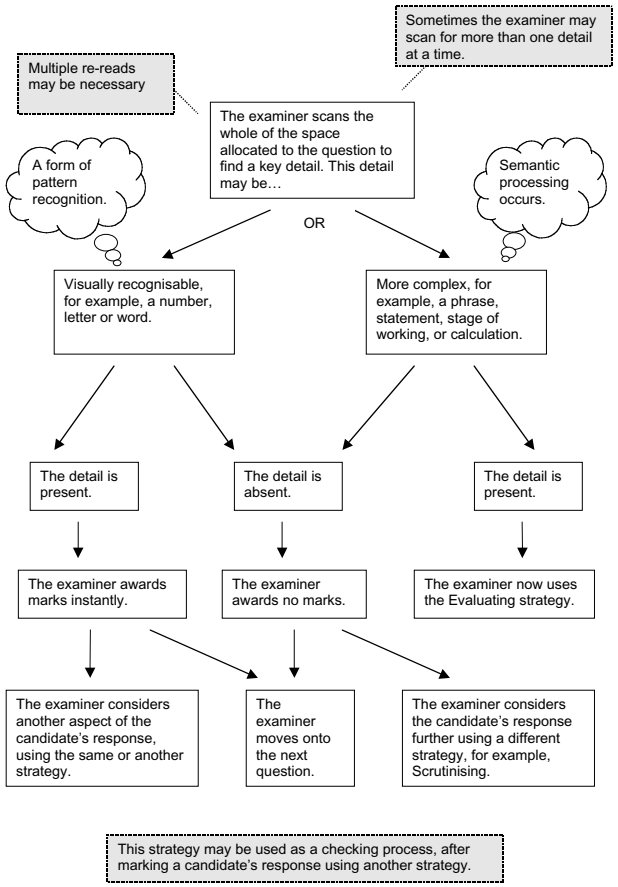


Figure 4: The 'Evaluating' strategy

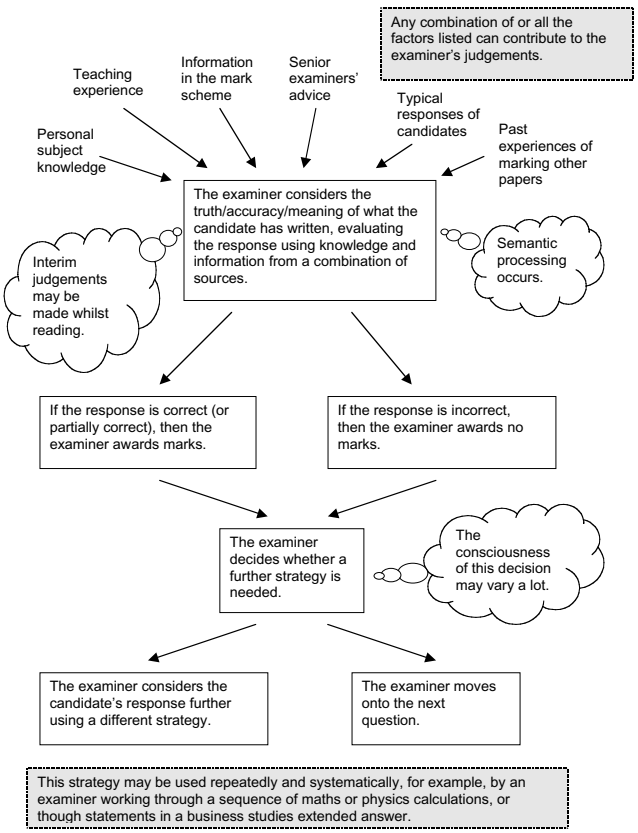


Figure 5: The 'Scrutinising' strategy

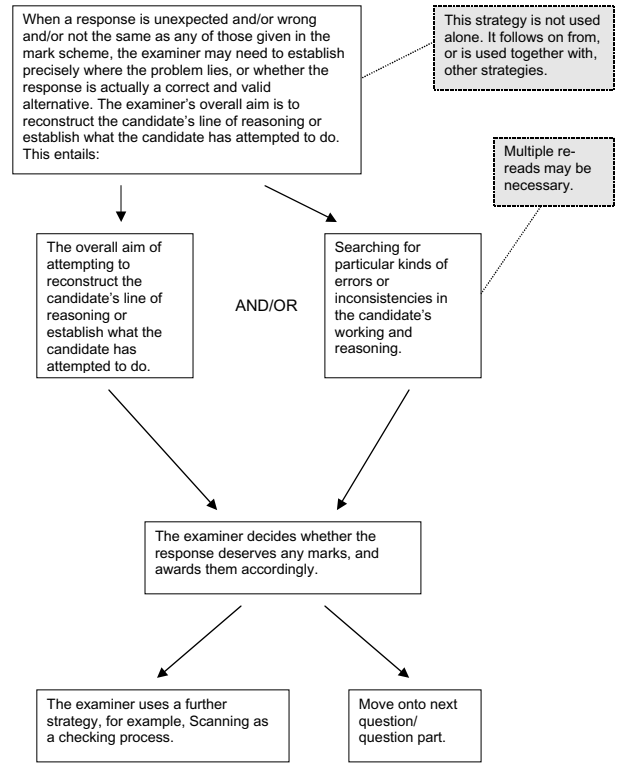
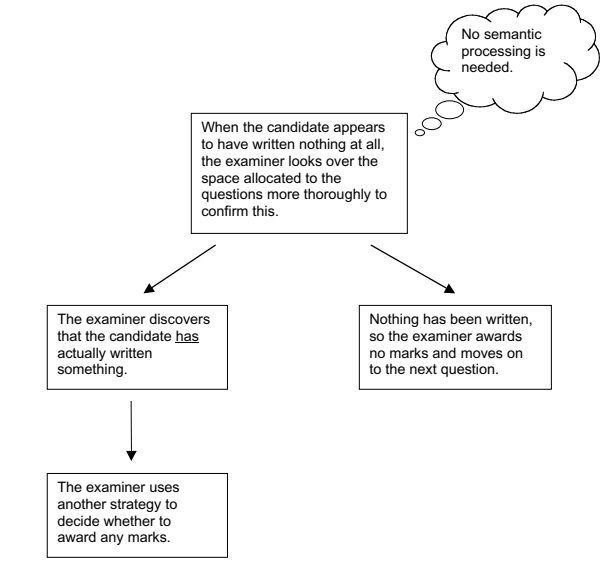
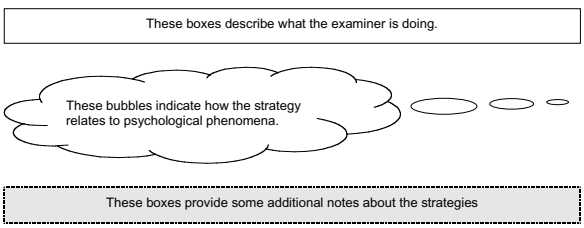


Figure 6: The 'No response' strategy



Key to Figures 1 to 6



Figures 7 and 8 contain transcript excerpts that can be taken to illustrate the five cognitive marking strategies. As these figures indicate, the marking strategies can be interpreted within dual-processing theories of judgement as comprising essentially System 1 judgement or System 2 judgement, or even both.

Figure 7: Examples of Mathematics examiners using each strategy

Verbal protocol excerpt	*Strategy used*	*System 1('intuitive') or System 2 ('reflective') judgements?*
Number seventeen: 61.49 instead of sixty-nine pounds seventy, so no marks there.	Matching	System 1
While I do that I'm sort of staring at the page, and I can see a four sixty-eight. Whoopee!	Scanning	System 1
We're looking for them using sine to find the angle.	Scanning	System 2
His algebra is fine.	Evaluating	Both Systems 1 and 2
Ugh, yuk. What a mess... Ah, looking at a whole load of percentages, all over the place. Er, and...that's a calculator paper. You wouldn't know it 'cause they're using non-calculator methods...and...that for division by a hundred. Can't...do it without doing a long division, poor dears. But it's all a load of... I'm trying to, erm, get close. It's all in trial and improvement... Erm, M0 for that one. Trying to, don't know how to divide things.	Scrutinising	Both Systems 1 and 2
And part E: no response so that gets nothing.	No response	System 1

Figure 8: Examples of Business Studies examiners using each strategy

Verbal protocol excerpt	*Strategy used*	*System 1 ('intuitive') or System 2 ('reflective') judgements?*
Four two five three is the answer. Four...no, no, nope, nope. No marks.	Matching	System 1
The answer on this one is 'Rea Aitkin, chairman', so as soon as I see 'Rea Aitkin, chairman', it's two marks.	Scanning	System 1
And looking for an action by Belgian chocolate manufacturers...	Scanning	System 2
The community, a judgement is made that the community should be considered, and a reason is that because they are consumers and obviously that would affect sales. Only a simple answer: one mark.	Evaluating	Both Systems 1 and 2
Now unusually, this candidate is suggesting, er, Miss Singh...as a decision-maker. I'm just checking the...this is the finance director. Erm, I'm accepting that, because a finance director can, sometimes, hold key decision-making influence. Er, I'm looking for a reason. 'He deals with the finances.' That's okay. I've accepted that for two marks.	Scrutinising	Both Systems 1 and 2
Blank: nothing.	No response	System 1

Discussion

The aims of this study of GCSE examiners were to identify the key marking strategies used, and to interpret them within the context of dual processing theories of judgement. There were several limitations, which included: the use of small samples of examiners; the exploration of just two GCSE examinations; four examiners not managing to mark all of their scripts; qualitative analysis inevitably involves some interpretation by the researchers and the potential of the process of 'thinking aloud' to interfere with the thought processes under investigation (for example, slowing them down). Together, these restrictions mean that our model of strategies is unlikely to be exhaustive.

Nevertheless, our study has some important implications. First, the complexity of some of the strategies identified confirms that GCSE examination marking can be a cognitively demanding process, often requiring considerable expertise. For some questions, the simpler strategies could, arguably, be used by many people, including those without much expertise and experience. However, those strategies that rely on subject knowledge, past marking and/or teaching experience, and on advice from the Principal Examiner, for example, *evaluating* and *scrutinising*, are often necessary when a candidate's response is long or unexpected.

Secondly, knowledge of the strategies identified in this study may prove useful to senior examiners. While several examiners have suggested that our named strategies provided a useful language with which to communicate with colleagues, others have suggested using the research in training courses for new examiners.

Thirdly, the study provides grounds upon which to hypothesise that some of the judgements entailed in marking may start off as slow and conscious System 2 thought processes, but migrate to System 1 as an examiner either acquires expertise or gains confidence. Examiners who were interviewed about the study supported this hypothesis, and several raised concerns about some examiners switching from using System 2 to using System 1 on particular questions before they were ready to do so. Several individuals felt that knowledge of the strategies would provide a means of 'self-checking' for all examiners, who could thereby remind themselves periodically of the need to evaluate and scrutinise some responses.

Finally, explicit knowledge of the strategies could prove useful when designing examination papers and mark schemes. For example, although it is impossible to predict every potential answer to a given question, listing as many valid responses as possible in the form of bullet points when the *matching* strategy is most likely to be used, or listing key information to scan for where a scanning strategy is viable, could help maximise the efficiency of the marking process.

Acknowledgements

We would like to thank Rita Nadas for preparing the diagrams for this article.

Further reading

A full report of the study described here is soon to be published in the *British Educational Research Journal* as an article entitled 'What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process'.

References

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, **7**, 31–51.

Ericsson, K. & Simon, H. (1993). *Protocol Analysis: Verbal reports as data*. London: MIT Press.

Kahneman, D. & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

Laming, D. (1990). The reliability of a certain university examination compared with the precision of absolute judgments. *Quarterly Journal of Experimental Psychology*, **42A**, 239–54.

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

Leighton J. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, Winter, 6–15.

Milanovic M., Saville, N. & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.). *Studies in Language Testing 3: Performance testing, cognition and assessment – Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem*. Cambridge: Cambridge University Press/University of Cambridge Local Examinations Syndicate.

Stanovich, K. & West, R. (2002). Individual differences in reasoning. In T. Gilovich, D. Griffin & D. Kahneman (Eds.). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

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

Van Someren, M., Barnard, Y. & Sandberg, J. (1994). *The think aloud method: A practical guide to modelling cognitive processes*. London: Academic Press.

Vaughan, C. (1992). Holistic assessment: what goes on in the rater's mind? In L. Hamp-Lyons (Ed.). *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex.

Webster, F., Pepper, D. & Jenkins, A. (2000). Assessing the undergraduate dissertation. *Assessment and Evaluation in Higher Education*, **25**, 71–80.

Yorke, M., Bridges, P. & Woolf, H. (2000). Mark distributions and marking practices in UK higher education. *Active Learning in Higher Education*, **1**, 7–27.

Examiners' annotations: Practice and purpose

Victoria Crisp and Martin Johnson Research Division

Introduction

> *'When you come to any passages that seem to you useful, make a firm mark against them, which may serve as lime in your memory, less otherwise they might fly away.'*
>
> Advice from St Augustine in Petrarch: *Secretum Meum* 1358

The processes of reading and writing are recognised to be inextricably intertwined. Writing helps to support cognitive demands made upon the reader whilst processing a text (e.g. O'Hara, 1996; Benson, 2001). Anderson and Armbruster (1982) suggest that annotating activities are concurrent with the actual reading processes, influence the way that reading occurs, and the way that meaning is processed. Examiners annotate scripts whilst marking (e.g. underlining, circling, using abbreviations or making comments) and this may reflect the cognitive support for comprehension building that annotations can provide.

Within the accountability agenda that pervades education there is an emphasis on clear communication channels between examiners of different seniority to facilitate effective monitoring. Annotations might have an important communicative role in this quality control process by offering others up and down the chain an insight into the rationale behind the annotating examiners' decisions. Previous re-marking investigations have suggested that annotations do have a communicative function, potentially influencing how subsequent viewers perceive the quality of a script (Murphy, 1979; Wilmut, 1984; Newton, 1996). Laming (2004) suggests that this is because there are places where the mark scheme leaves the examiner uncertain, and that judgements in such cases are influenced by extraneous information, for example, the previous annotations of other judges.

In addition to evidence that annotations act as a communicative device, there is also evidence that annotating might have a positive influence on markers' perceptions and affect their feelings of efficacy. Most markers felt that annotating improved their marking, helping them to apply performance criteria and reducing the subjectivity of judgements (Bramley and Pollitt, 1996). In pilot work on online assessment teachers, examiners and moderators have expressed dissatisfaction where facilities for annotation were limiting (Greatorex, 2004; Raikes *et al.*, 2004). Markers report that using annotations provides an efficient means to confirm or reconsider standards both within and across candidates as well as acting as a reassurance during the judgemental process (Shaw, 2005).

Rationale

The literature available provides some information about the purposes and effects of annotations. However, there is a relative sparsity of published research about annotation in examination marking in terms of the following:

- consistency of use of codes
- examiners' reasons for using annotations
- the role that annotations might be playing in decision making processes
- the effects, or perceived effects, of using annotations whilst conducting first marking.

This research investigates some of these issues and develops a more comprehensive picture of annotation practices.

Annotations in different subjects

Before the main study, we analysed some marked examination scripts to build a basic profile of annotation use across a variety of subject disciplines. Three scripts from each of sixteen subjects were analysed. The types of annotation used and the frequency of their use varied substantially between subjects. However, six types of annotation were used fairly frequently across a number of subjects:

- making comments
- underlining
- sidelining
- circling
- question marks
- carets.

There were no significant relationships between frequency of annotation and subject families. However, analysis did show a significant relationship between the length of response and mean number of annotations ($p = 0.007$) with comments, in particular, being more common where questions require longer answers.

Main study method

The main study investigated one Mathematics GCSE and one Business Studies GCSE paper. Six examiners (including the Principal Examiner) who had previously been standardised to mark the paper were recruited for each subject.

Examiners initially marked ten scripts which were then reviewed by their Team Leader. Examiners then marked a further 46 (Business Studies) or 40 (Mathematics) scripts. A representative sample of scripts was used. The scripts used were photocopies of the original scripts with all marks and annotations removed to simulate a first marking situation. Each examiner within a subject marked the same candidates' scripts. The reliability of the experimental marking was investigated and analyses suggested that marker behaviour in this study was generally representative of marker behaviour during live marking (see Crisp and Johnson, *in press*, for details).

The examiners later attended individual meetings with researchers. The session began with each examiner marking a small number of new scripts to re-familiarise themselves with the examination paper and mark scheme. A researcher then observed each examiner as they continued marking a few further scripts. This recorded whether individual annotations were made before or after the recording of marks (for details and findings of this stage see Crisp and Johnson, *in press*). Later in the session each examiner was interviewed about their use of annotations.

Analysis of annotation practices

Annotation use in the two subjects differed in a number of ways (see tables below). Business Studies markers used roughly twice as many annotations per question part as Mathematics markers. Furthermore, the most common types of annotations used were different for the two subjects. However, in both subjects markers varied in the frequency and types of annotations they used compared with others in their team, with no obvious relationship found between marker reliability and annotation use between markers. For Business Studies there were significant differences between the frequencies of use of most annotations across examiners, perhaps reflecting different preferences or habits. For Mathematics, despite sometimes significant variations between markers in frequency of annotation use, there was an underlying similarity in the types of annotations used.

Mean number of annotations per script (Business Studies)

Examiner	*BOD*	*Underline*	*Comment*	*Sideline*	*Level*	*Question Mark*	*All annotations*
1	0.83	0.00	0.00	0.00	0.00	0.00	0.91
2	0.87	2.96	2.78	2.61	3.43	3.00	17.52
3	1.70	1.61	0.48	1.30	8.52	0.13	14.96
4	0.43	3.96	5.22	0.09	5.96	0.91	17.87
5	0.04	0.65	0.39	1.70	11.83	0.22	15.87
6	0.35	6.39	0.52	0.26	9.52	0.17	17.48
All examiners	0.70	2.59	1.57	0.99	6.54	0.74	14.10

Mean number of annotations per script (Mathematics)

Examiner	*Circle*	*Underline*	*Comment*	*Caret*	*M*	*Sideline*	*All annotations*
1	1.35	0.95	0.55	1.60	1.15	0.00	6.70
2	1.00	1.05	0.70	2.45	2.50	0.50	10.00
3	2.15	1.35	0.95	1.90	2.40	1.53	12.26
4	1.05	1.10	0.10	3.30	1.75	0.45	8.95
5	1.85	0.55	0.95	2.00	1.75	0.65	8.85
6	1.40	4.05	0.95	2.40	2.25	0.15	13.00
All examiners	1.47	1.51	0.70	2.28	1.97	0.54	9.94

Note: Some of the less frequently used annotation types have been omitted from these tables

In both subjects some markers sometimes employed their own stylistic annotations in preference to commonly recognised standard annotations used by others. For example, instead of using circles to indicate errors one marker chose to place errors in parentheses.

Analysis of interviews

Interviews were carried out to probe examiners' ideas about the purposes of annotating, where examiners gained their annotation knowledge, and to elicit their perceptions of annotating.

Using annotations to justify decisions to others appeared to be the most salient purpose for all Mathematics markers. The Business Studies examiners also reported justifying or explaining decisions to be one of the purposes of annotation and placed additional emphasis on communicating to other examiners the reasoning behind awarding marks. One examiner said, *'It's a way of telling somebody else why and when I reach a decision. Er, whether they agree with it or not, at least they can see where I'm coming from.'*

Examiners also reported that annotating supported their judgements (particularly where marginal decisions had to be made) and helped them reach the appropriate mark. For example, one examiner said, '*I feel it's there to help the examiner reach the correct mark for the question; so it's an aid*'. Another said, '*So my annotation really is my way of thinking aloud but without talking aloud. Er, it's my way of communicating to myself.'*

Most markers felt that annotating helped to structure their thinking whilst marking, particularly for questions where candidates needed to

show their working or where multiple marks were available in Mathematics or when using levels of response marking in Business Studies.

All examiners reported that the annotations they used came from guidance in the mark scheme or (in Business Studies) from an information sheet provided separately. These were often used to support brief discussion at standardisation meetings. Annotation conventions were often reported to be well established and commonly understood among markers, with relatively little room for markers to use non-standard annotations. This was especially the case amongst Mathematics examiners. Half of the Mathematics markers suggested that the community's annotation conventions were based *'pretty well in folklore'*, *'word of mouth'* or *'custom'*. Other markers conveyed a profile of a centralised marking community with clear guidance about how to apply conventions. Two Business Studies examiners reported that their annotations were based on what the Team Leader and other examiners used. Those Business Studies examiners who gave a view thought that the types of annotations used were fairly consistent within the examining team although patterns and frequency of use might vary.

However, in both subjects there was space for marker idiosyncrasies provided that it did not compromise common understanding. Some examiners explained that they felt it necessary to sometimes use full words and sentences rather than standard annotations in order to communicate their reasoning clearly.

All but one marker felt that annotating was a positive aspect of marking. Again, the dual functions of accountability (*'It's useful for the examiner, the chief examiner, to know how you've allocated the marks'*) and supporting marker judgements (*It gives more clarity in your own mind when you're awarding marks'*) were the most clearly stated reasons used to support this. The examiner who viewed annotation negatively expressed two reasons for this opinion: first that he/she was not *'fully aware of how to use them to their best effect'* and secondly that he/she did not see the need to use them saying that the *'mark scheme is what I'm marking to'*.

Discussion

The findings portray a clear sense that markers in both subjects believed that annotating performed two distinct functions. The first appeared to be justificatory, communicating the reasons for their marking decisions to others. This mirrors the statutory requirements for awarding bodies to establish transparent, accountable procedures which ensure quality, consistency, accuracy and fairness. The second purpose was to support their thinking and marking decisions. In addition to helping markers with administrative aspects of marking (for example, keeping a running tally of marks) there are claims that annotations also support higher order reading comprehension processes.

There are also suggestions that annotations can help to provide a 'visual map' of the quality of answers (Bramley and Pollitt, 1996). This is perhaps especially useful for the purpose of making comparisons – especially between longer texts that possibly exact a great deal of cognitive demand on the marker. Laming (2004) suggests that making comparisons is a key element in the process of making judgements. So it is not surprising that the notion of 'annotating to support thinking' appeared to be more salient for Business Studies markers, who were more likely to deal with longer answers, than for Mathematics markers.

From the study it appears that different subjects have different annotation profiles. There seems to be a 'pool' of annotations pertinent to each subject, and structures exist within subject marking communities to transmit information about appropriate annotation use, for example, through mark schemes and standardisation meetings.

This said, individual markers dipped into their subject 'annotation pool' in different ways. It was common for markers to have their own particular annotation profile, using different subsets of available annotations and using certain annotations with varying frequency. It was uncommon for examiners to refer to annotation lists whilst marking, suggesting that their annotations were internalised. Preston and Shackelford (1999) report similar findings in the context of assessing computer science. They found that raters did not refer to the list of feedback codes used to classify errors but that 'raters remember and use a small subset of all available (and appropriate) feedback codes while marking' (p. 31). This variation between examiners could seem negative but it is clear that the use of annotation is currently an efficient communicative practice that is an automatic part of the marking process for those embedded in the examining culture, and rarely something that is 'tagged on'. Examiners may vary in their usage but whatever they are doing it appears to support their work and they are positive about the role that it plays.

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. This reflects the role of annotation as a communicative tool, reflecting notions embedded in Situated Learning Theory. This 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. Part of this communication might involve the evolution and use of specialised tools that 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. Study findings suggest that markers believe annotating to be a positive aspect of marking which concurs with other findings (Bramley & Pollitt, 1996).

This research has gathered evidence on current annotation practices in examination marking in two subjects. Whilst the data available do not suggest that patterns of annotation use dramatically affect marker reliability, the practice of annotation sometimes supports marker judgement, is generally viewed positively and appears to give examiners confidence in their professional judgements.

Acknowledgements

We would like to thank the examiners involved for their participation and for sharing their views with us. We would also like to thank Jackie Greatorex and John F. Bell for the statistical analysis relating to the reliability of the experimental marking.

Further reading

This report is a summary of an article, 'The use of annotations in examination marking: opening a window into markers' minds', which is *in press* at the *British Educational Research Journal*.

References

Anderson, T. H. & Armbruster, B. B. (1982). Reader and text-studying strategies. In W. Otto & S. White (Eds), *Reading Expository Material*. London: Academic Press.

Benson, P. J. (2001). Paper is still with us. *The Journal of Electronic Publishing*, **7**, 2. Available online at: www.press.umich.edu/jep/07-02/benson0702.html (accessed 25 August 2005).

Bramley, T. & Pollitt, A. (1996). *Key Stage 3 English: Annotations Study*. A report by the University of Cambridge Local Examinations Syndicate for the Qualifications and Curriculum Authority. London: QCA.

Crisp, V. & Johnson, M. (*in press*). The use of annotations in examination marking: Opening a window into markers' minds. *British Educational Research Journal*.

Greatorex, J (2004). Moderated e-portfolio project evaluation. Evaluation and Validation Unit, UCLES. Available online at: www.ocr.org.uk/OCR/WebSite/Data/Publication/E-Assessment%20Materials/Moderated_82372.pdf (accessed 25 August 2005).

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

Murphy, R. (1979). Removing the marks from examination scripts before remarking them: Does it make any difference? *British Journal of Educational Psychology*, **49**, 73–8.

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.

O'Hara, K. (1996). *Towards a typology of reading goals: RXRC affordances of paper project*. Technical Report EPC-1996-107. Available online at: www.lergonome.org/pdf/EPC-1996-107.pdf (accessed 25 August 2005).

Preston, J. A. & Shackelford, R. (1999). Improving on-line assessment: An investigation of existing marking methodologies. *ACM SIGCSE Bulletin,* **31**, 3, 29–32.

Raikes, N., Greatorex, J. & Shaw, S. (2004). *From paper to screen: some issues on the way*. Paper presented at the International Association of Educational Assessment conference, Philadelphia, June. Available at: www.ucles.org.uk/assessmentdirectorate/articles/confproceedingsetc/IAEA2000NRJGSS (accessed 25 August 2005).

Shaw, S. (2005). *On-screen marking: investigating the examiners' experience through verbal protocol analysis*. University of Cambridge ESOL Examinations, Report No 561.

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 RAC 315.

Judging learners' work on screen: Issues of validity

Martin Johnson and Dr Jackie Greatorex Research Division

Introduction

Current developments in Cambridge Assessment and elsewhere include assessors[1] marking digital images of examination scripts on computer, rather than the original scripts on paper, and judges marking and moderating digitally produced coursework on computer, rather than on paper. One question such innovations raise is whether marks from judgements made about the same work presented on computer and on paper are comparable.

Generally the literature concerning the on-screen marking of tests and examinations suggests that on-paper and on-screen scores are indeed comparable (e.g. Bennett, 2003; Greatorex, 2004), although Fowles and Adams (2005) report that differences have been found in studies by Whetton and Newton (2002), Sturman and Kispal (2003) and Royal-Dawson (2003).

Our concern in this discussion article is that even when double marking studies find high levels of agreement between marks for the same work judged in different modes, issues of validity might be masked. We are thinking of validity in terms of the cognitive processes of the assessor when reaching a judgement, judgement per se and how well these reflect the judgements that were intended when the assessment was devised.

1. We use the terms assessor/judge as general terms to refer to examiners, moderators, verifiers or others who assess candidates' work.

Do assessors all use the same criteria/guidelines in the same way?

If assessors do not use the same criteria/guidelines in the same way then validity is threatened. Laming (2004) argues that all judgements are comparisons of one thing with another and this process is influenced by the prior experience of the judge and the context in which the comparisons are being made. In the case of examination marking he explains that sometimes the mark scheme might leave room for examiner uncertainty, especially when marking essays, when mark schemes might be interpreted in different ways.

Other research evidence has suggested that assessors do not mechanistically match learners' achievement to assessment criteria (Wolf, 1995) and that a range of extraneous factors can influence assessors' decisions (Ragat and Hevey, 1995). These might include a respondent's handwriting (Green *et al*., 2003; Milanovic *et al*., 1996, Sanderson, 2001), the context in which an assessment is carried out (Wolf, 1995), or the assessor's own idiosyncratic internalised standards (Eraut *et al*., 1996).

Although the above research review is partial, it is intended to illustrate that human judgement in conventional assessment practices is potentially influenced by a number of extraneous variables, partly explaining why assessment criteria might be interpreted by different judges in different ways. We would like to explore how the mode of marking/moderating might also lead to the influence of extraneous variables.

If assessors are presented with the same work in different modes do they make qualitatively different judgements?

We are particularly interested in the area of assessment that is likely to place a high cognitive demand on the assessor, for example, the assessment of extended prose. It is possible that presenting such work on screen might affect how assessors read and perceive the text, and therefore, possibly, their assessment judgements.

The concept of affordance recognises that the environment influences subjects' behaviour, with some environments facilitating or inhibiting certain types of activity. Gibson (1979) suggested that part of the success of human evolutionary development has been a consequence of their ability to identify and exploit the affordances of different environments. He claimed that humans perceive affordance properties of the environment in a direct and immediate way and they subsequently perceive possibilities for action. In this sense, the modes of paper and computer exist as environments within which activity is carried out, and each has its own affordances.

An interesting implication of Gibson's theory concerns the evolution of behaviour as a response to environmental change. Using the Piagetian notion of learning as a function of assimilation and accommodation, Reinking *et al*. (2000) suggest that a lag may exist between technological developments and the accommodation of those changes as reflected in the behaviour, and specifically the reading and writing activities, of people interacting with that new technology. This implies that patterns of behaviour prevalent when working in a paper-based environment may continue even though the terrain of that environment has been changed by new computer technology.

Sellen and Harper (2002) compare the affordances of paper and of digital reading technologies.

Paper	*Digital reading technologies*
• Able to support flexible navigation	• Able to store and access large amounts of information
• Able to support cross document use	• Able to display multimedia documents
• Able to support annotation while reading	• Enable fast full-text searching
• Able to support the interweaving of reading and writing	• Allow quick links to related materials
	• Allow content to be dynamically updated or modified

Sellen and Harper (2005) suggest these affordances are linked to the purposes of reading and text length. Where the focus of reading only requires a superficial skim of a longer text, or the deeper processing of shorter text (i.e. text that can be confined to one screen at a reasonable resolution), mode-related effects on reading might be minimal.

A number of studies have investigated mode-related marking effects. Sturman and Kispal (2003) studied differences in Key Stage 2 spelling, reading and writing marks given when the same work was marked by examiners using a computer package to view the work and record the score and when it was marked on paper. Although there was an absence of any consistent trend, the differences between mean marks might suggest e-marking and conventional marking judgements were qualitatively different. Findings from studies by Price and Petre (1997) and Greatorex (2004) have identified a number of mode-related aspects that might qualitatively influence assessment judgements.

Page and document management

Sellen and Harper (2002) suggest that digital technologies have significant constraints when dealing with document layout, largely because any document needs to be presented within a particular screen size, which limits how many pages or documents can be made visually available at once. They contrast this with page and document management techniques on paper, where navigation is two-handed and allows the simultaneous overlap and interleafing of pages and documents. These techniques allow pages to be manoeuvred so that visual connections can be made, affording the reader a broader field of view than is possible on screen. Both Greatorex (2004) and Price and Petre (1997) find evidence to suggest that mode might affect judgement processes. In the Greatorex study moderators suggested that reading and navigating through e-portfolios was hindered because scrolling backwards and forwards was difficult. Viewing different candidates' work was another source of difficulty. In some situations moderators intended to almost simultaneously search for evidence in different pieces of work by the same candidate, but could not do this on screen as there seemed much more of a time lag between reading one piece and another. Where candidates had used more than one file to present their work, dipping in and out of multiple files to identify where the teacher had given credit was also burdensome and time-consuming. Price and Petre also reported that markers in their study found opening and switching between documents to be onerous.

Technological mediation of the text

Technology might draw an assessor's attention to information to which they would not otherwise have attended. Greatorex reports that some moderators noted that many of the e-portfolio files were in Microsoft Word™ which incorporated an automatic spell and grammar check, underlining spelling and possible grammar errors in candidates' work. These moderators suggested that this software made it easier to see Quality of Written Communication (QWC) errors on screen compared with the paper version of the same portfolios. The process involved in making the judgement about the e-portfolio and the paper version of the same portfolio were qualitatively different. This is an important issue since it reiterates the crucial relationship between assessment tools, assessment purpose and underlying validity.

Sense of text

Greatorex (2004) reported moderators trying to find their way around candidates' work to gain a sense of meaning from the text. E-moderation required moderators to scroll to see all the evidence for an assessment criterion, affording them the opportunity to see the information in snapshots with a limited view of the whole text. In contrast, when using paper portfolios they could glance from one area to another whilst maintaining a view of the broader text.

A wealth of research exists to suggest that spatial encoding takes place during the reading process and that this is an integral part of building a mental representation of the location of textual information (Piolat *et al*., 1997; Fischer, 1999; Kennedy, 1992). Piolat *et al*. cite a number of studies

which suggest that paper supports the process where readers assign geographical locators to words and ideas during reading. It is inferred that this is not equally afforded by both paper and computer-based texts, in turn having implications for a reader's cognitive load. Pommerich (2004) has found evidence to suggest that readers' positional memory is better on paper because it appears that they can operationalise positional memory more easily. Reasons for this appear to relate to the fact that only having a limited view of a text on screen disturbs the reader's representation of its spatial layout and disrupts their construction of a reliable mental representation of the text. Research by Dyson and Haselgrove (2000) has also found evidence of subjects' extremely poor on-screen reading performance on structural questions that require the locational recall of at least two pieces of information. One explanation for this relates to differences in navigability around texts. Pommerich suggests that the scrolling techniques afforded by computer-based texts only allow relative spatial orientation since the field of vision is less than when navigating around paper-based texts. Furthermore, the relative slowness and imprecision of scrolling compared with manual navigation may lead to increased cognitive demands on the reader whilst they find their way around the text since this 'does not provide…enough "tangible" data about the location of information that is not currently on the screen. Each time the scroll arrows are used, or even the scroll bar, the spatial layout is disrupted', and moreover, 'scrolling through the text to find a particular piece of information can be slow' (Piolat *et al*., p.568). To conclude, Piolat *et al*. argue that the combination of slow and imprecise navigation around a text, disrupted spatial layout and the limited view of the text on screen make it difficult for a 'sense of text' to be constructed when accessed on computer.

Such findings imply that it is more cognitively demanding to gain a sense of text when e-moderating or e-marking than when assessing on paper.

Reading strategies

Greatorex (2004) reported that mode somewhat influenced how moderators read candidates' work. Teachers and moderators search through the portfolio to look for particular information to satisfy the assessment criteria. Some moderators reported that in reading paper portfolios they had spotted evidence that appeared to have been missed by teachers and moderators who previously assessed the electronic version of the same portfolios on screen.

O'Hara (1996) described a series of reading strategies. He found that (1) reading strategy choices are related to the purpose of the reading, for example, proof reading requires different strategies to reading for information, and (2) mode has a greater influence on reading strategies than the purpose of the reading. Askwall (1985) described a number of search strategies used by readers and showed that search strategy choices were influenced by mode. The results of Greatorex (2004) (mentioned above) are in keeping with this research literature. Therefore, the information gleaned by assessors about candidates' work and the sense they make of it might be affected by mode.

Annotation

Price and Petre (1997) found that assessors in their study had different marking styles on paper and on screen, but that some markers were more affected by mode than others. They also found that annotations used in paper and e-marking were different despite being available in both modes. Greatorex found that when teachers marked on screen they reported difficulties annotating directly onto candidates' work and said there would have been more annotations if the portfolios had been on paper.

Annotating can help text comprehension and affects how meaning is processed (O'Hara, 1996). In some circumstances it supports markers' decision-making processes (Crisp and Johnson, *in press*). Making paper-based annotations is relatively effortless and is part of the meaning construction process during reading but computer-based annotation might be impeded by a lack of usable annotation tools (O'Hara and Sellen, 1997). A number of recent developments have been designed to overcome these weaknesses, including stylus entry directly onto a tablet or touch screen, and digital ink technologies.

We deduce from the research literature that annotation plays a crucial role in text comprehension and that in some situations this might be important when making assessment judgements.

Conclusions

Although the above research review is partial, it is intended to illustrate that judgements are potentially influenced by a number of extraneous variables. There is some evidence that mode does not generally affect scores across different modes but that judgements are sometimes affected qualitatively, in which case validity can be enhanced or compromised.

If it is the case that mode affects assessment judgements, it must be considered in the wider context. First, it is just one of many influences, and secondly, the benefits of technology in education and assessment are well rehearsed in a large body of research literature (e.g. Salmon, 2004; Heppell, 2003; Sellen and Harper, 2002).

References

Askwall, S. (1985). Computer supported reading vs. reading text on paper: a comparison of two reading situations. *International Journal of Man Machine Studies*, **22**, 425–439.

Bennett, R. E. (2003, November). *On-line assessment and the comparability of score meaning* (ETS Publication, RM-03-05). Retrieved August 31, 2005 from http://ftp.ets.org/pub/res/researcher/RM-03-05-Bennett.pdf

Crisp, V. and Johnson, M. (*in press*). The use of annotations in examination marking: opening a window into markers' minds. *British Educational Research Journal*.

Dyson, M. C., and Haselgrove, M. (2000). The effects of reading speed and reading patterns on the understanding of text read from screen. *Journal of Research in Reading*, **23**, 2, 210–23.

Eraut, M., Steadman, S., Trill, J. and Parkes, J. (1996, November). *The assessment of NVQs*. (Research Report No. 4) Brighton: University of Sussex Institute of Education.

Fischer, M. H. (1999). Memory for word locations in reading. *Memory*, **7**, 1, 79–118.

Fowles, D. and Adams, C. (2005). *How does assessment differ when e-marking replaces paper-based marking?* Paper presented at the IAEA Conference, Abuja, Nigeria. Retrieved February 5, 2006 from www.iaea.info/abstract_files/paper_051218101528.doc

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Greatorex, J. (2004, December). *Moderated e-portfolio project evaluation*. (Evaluation and Validation Unit, University of Cambridge Local Examinations Syndicate). Retrieved August 31, 2005 from http://www.ocr.org.uk/OCR/WebSite/Data/Publication/E-Assessment%20Materials/Moderated_82372.pdf

Green, S., Johnson, M., O'Donovan, N. and Sutton, P. (2003). *Changes in key stage two writing from 1995 to 2002*. Paper presented at the United Kingdom Reading Association Conference, University of Cambridge, UK, July 2003.

Heppell, S. (2003). Assessment and new technology: New straitjackets or new opportunities. In C. Richardson, (Ed.), *Whither Assessment?* London: Qualifications and Curriculum Authority. Retrieved August 31 2005 from http://www.qca.org.uk/downloads/combined_whither_assessment.pdf

Kennedy, A. (1992). The spatial coding hypothesis. In K. Rayner (Ed.), *Eye movements and visual cognition*. New York: Springer-Verlag.

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

Milanovic, M., Saville, N. and Shuhong, S. (1996). A study of the decision-making behaviours of composition markers. In M. Milanovic and N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge*. Studies in Language Testing 3. Cambridge, UK: Cambridge University Press/UCLES.

O'Hara, K. (1996). *Towards a typology of reading goals*. (Rank Xerox Research Centre Affordances of Paper Project Technical Report EPC-1996-107). Cambridge, UK: Rank Xerox Research Centre.

O'Hara, K. and Sellen, A. (1997). A comparison of reading paper and online documents. *Proceedings of the conference on human factors in computing systems (CHI '97)*, 335–342. New York: Association for Computing Machinery.

Piolat, A., Roussey, J-Y. and Thunin, O. (1997). Effects of screen presentation on text reading and revising. *International Journal of Human-Computer Studies*, **47**, 565–89.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based texts. *The Journal of Technology, Learning and Assessment*, **2**, 6. Retrieved August 31, 2005 from http://www.bc.edu/research/intasc/jtla/journal/pdf/v2n6_jtla.pdf

Price, B. and Petre, M. (1997). *Teaching programming through paperless assignments: An empirical evaluation of instructor feedback*. Centre for Informatics Education Research, Computing Department, Open University, UK. Retrieved April 20, 2005 from http://Mcs.open.ac.uk/computing papers/mzx/teaching.doc

Ragat, P. and Hevey, D. (1995). *Sufficiency of evidence: The development of guidance for Assessors and Verifiers*. (Research and Development Report No. 32), Sheffield, UK: DFEE.

Reinking, D., Labbo, L. D. & McKenna, M. C. (2000). From assimilation to accommodation: A developmental framework for integrating digital technologies into literacy research and instruction. *Journal of Research in Reading*, **23**, 2, 110–22.

Royal-Dawson, L. (2003). *Electronic marking with ETS software*. AQA Research Committee. Paper RC/219. In D. Fowles and C. Adams (2005), *How does assessment differ when e-marking replaces paper-based marking?* Paper presented at the IAEA Conference Abuja, Nigeria. Retrieved February 5, 2006 from www.iaea.info/abstract_files/paper_051218101528.doc

Salmon, G. (2004). *E-moderating*. The key to teaching and learning on-line. London, UK: Routledge Falmer.

Sanderson, P. J. (2001). Language and differentiation in examining at A level. PhD thesis, University of Leeds.

Sellen, A. and Harper, R. (2005). Personal correspondence.

Sellen, A. and Harper, R. (2002). *The myth of the paperless office*. Cambridge, MA: MIT Press.

Sturman, L. and Kispal, A. (2003). *To e or not to e? A comparison of electronic marking and paper-based marking*. Paper presented at the 29th International Association for Educational Assessment Annual Conference, Manchester, UK, October 2003.

Whetton, C. and 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 2002.

Wolf, A. (1995). *Competence-based assessment*. Buckingham, UK: Open University Press.

The Cambridge Assessment/Oxford University automatic marking system: Does it work?

Nicholas Raikes Research Division

In the first issue of *Research Matters*, Sukkarieh *et al*. (2005) introduced our work investigating the automatic marking of short, free text answers to examination questions. In this article I give details and results of an evaluation of the final prototype automatic marking system that was developed.

Introduction

Background

Cambridge Assessment funded a three year research project that investigated the application of computational linguistics techniques to the automatic marking of short, free text answers to examination questions. The research, conducted at Oxford University by Professor Stephen G. Pulman and Dr. Jana Z. Sukkarieh, focussed on GCSE Biology as a suitable context since the Biology question papers contained large numbers of questions requiring short, factual, written answers.

The researchers took two broad approaches to automatic marking. The first approach involved writing by hand what were, loosely speaking, machine marking schemes for the items to be automatically marked. This approach is referred to as the 'Information Extraction' approach. The second approach – dubbed the 'Machine Learning' approach – involved trying various machine learning techniques to, again loosely speaking, learn the marking scheme from a sample of human marked answers. A hybrid approach using semi-automatic methods to produce

the machine marking scheme was also investigated.

Much useful research with some promising results was done in relation to the machine learning and hybrid approaches, offering the prospect of reducing the amount of specialised work required to set up new items. For details, see Pulman and Sukkarieh (2005). A complete prototype marking system was developed using Information Extraction techniques, and it is this system that is the focus of the evaluation reported in this article.

How the system works

We gave information about the system in our last *Research Matters* article (Sukkarieh *et al*., 2005). In essence the system matches answers to be marked against pre-written patterns to extract pertinent information previously judged by human examiners to warrant the award or forfeiture of a mark. The patterns can include syntactic information to specify parts of speech, verb groups and noun phrases, and essentially a pattern covers the synonyms for each pertinent piece of information.

Patterns are written by hand and based on the marking scheme used by human examiners, together with a sample of 200 human-marked answers. The sample answers are annotated by the human examiners to indicate precisely the part(s) of each answer which gained or forfeited marks – this annotation is done to minimise the need for the person writing the patterns to make these judgements.

Method

Two multi-part Biology questions were chosen from a 2003 GCSE Double Science examination. They were chosen because:

- They were common to both Foundation and Higher tiers and therefore could be used with the widest range of candidates.
- Every sub-part required a short, factual, textual, answer. This means that the whole questions could be used, providing a cohesive mini computer based test that can subsequently be given to volunteers to demonstrate or further research the system.
- There were eight 1-mark items (sub-parts) and five 2-mark items and so the questions covered the range for which this automatic marking technique might be suitable.

In the real GCSE examination, candidates answered on paper. Since automatic marking requires machine readable text, a random sample of 748 paper scripts was obtained and the relevant answers – excluding the totally blanks! – keyed into a computer file. Two hundred of these answers to each item were used to help with writing the patterns, while the remaining answers were held back for use in the marking trial.

All answers for each item had been live marked (i.e. marked for real in the GCSE examination) by human examiners; we had the resulting item-level marks keyed into a database. For the evaluation we recruited two senior examiners – both of whom had led teams of examiners in the live marking – to independently mark the transcribed answers a further two times. These examiners marked hard copies of the transcriptions; their marks were also keyed into the database. We therefore had three human examiner marks for each non-blank answer: one live[1] and one from each of the two Team Leaders recruited for this evaluation. In addition to marking the answers, we asked the two Team Leaders to annotate 200 of the answers to each item to show, by highlighting and labelling, precisely which parts of the answer matched each numbered point in the examiners' written marking scheme. The two hundred answers to each item were chosen according to their live marks as follows. For the 1-mark items, random samples of 50 0-mark answers and 150 1-mark answers were drawn. For the 2-mark items, the proportions were 50 0-mark, 75 1-mark and 75 2-mark. Where there was a shortage of higher mark answers, half of those available were used in the training data (the balance made up of lower scoring answers), and half were retained for use in trial marking.

The researchers in Oxford were provided with the following material to help them write the patterns:

- copies of the question paper and of the examiners' written marking scheme;
- both sets of the 200 annotated sample answers for each item;
- all three sets of marks for these answers (one live mark and two evaluation marks).

The Oxford researchers were **not** provided with any details of the remaining answers used for trial marking. These answers are referred to as the 'unseen answers'.

Oxford's patterns were sent to Cambridge Assessment and compiled into the automatic marking system running on a Cambridge Assessment server for trialling.

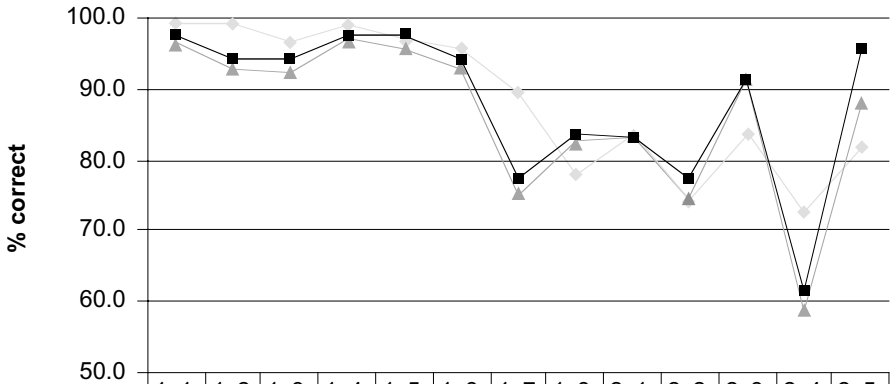
The unseen answers were marked automatically using the patterns developed by Oxford. The output included details of words that were unrecognised by the system, generally due to spelling errors made by candidates, together with suggested alternatives. If the test had been taken by candidates on computer, the system would have used this output to provide a spelling checking facility to candidates. It was therefore decided to correct the errors and run this corrected data through the marking engine. In this way the best and worst case scenarios for spelling mistakes could be compared.

In addition to the Oxford patterns, Cambridge Assessment also commissioned a temporary worker previously totally unacquainted with the project to write patterns for three of the items. This person had a background in psychology and computer science, but no experience of computational linguistics. He relied primarily on Oxford's documentation to train himself in how to write the patterns, and when writing his patterns had access to exactly the same material as Oxford had had. These patterns – the Cambridge Assessment patterns – were compiled into the system and the unseen answers were marked using them. In this way 'a proof of concept' investigation was conducted into the feasibility of transferring pattern-writing skills to persons not involved in developing the system – a key requirement for commercial use of the system.

Results

We will present and comment on the results in the following order. First we report on the correctness of the automatic marks using the Oxford patterns. Next we report inter-marker agreement levels, comparing each of the four markers (one automatic and three human) with each other; again, the Oxford patterns were used by the automatic marker. Finally, we report similar results for when the Cambridge Assessment patterns were used, and compare them with the previous Oxford results.

1. The live marks were not all due to a single examiner, since sample scripts were chosen at random.



	1_1	1_2	1_3	1_4	1_5	1_6	1_7	1_8	2_1	2_2	2_3	2_4	2_5
spelling corrected	97.5	94.2	94.4	97.9	97.6	94.5	77.3	83.9	83.2	77.4	91.2	61.4	96.0
spelling uncorrected	96.7	92.9	92.4	97.2	95.9	93.0	75.1	82.5	83.4	74.6	91.2	58.7	88.2
200 sample	99.5	99.5	96.5	99.0	97.0	96.0	89.7	78.0	83.5	74.5	84.0	72.5	82.0

items

Figure 1: Percentages of answers marked correctly by the automatic system

The automatic marks' correctness: Oxford patterns

Figure 1 shows, for the eight 1-mark items (1_1 to 1_8) and five 2-mark items (2_1 to 2_5):

- **Spelling corrected:** The percentage[2] of the **unseen** answers marked correctly by the system, after the answers were corrected for spelling.
- **Spelling uncorrected:** The percentage of the **unseen** answers marked correctly by the system, with no spelling corrections.
- **200 Sample:** The percentage of the **200 sample answers** to each item, used by Oxford for pattern writing, that the system marked correctly[3].

Table 1 presents the percentage of answers-per-item correctly marked by the system, averaged separately across the 1-mark and 2-mark items. The 1-mark average was also calculated with the two outlying items – 1_7 and 1_8 – excluded.

Key points concerning Figure 1 and Table 1 are:

- The system's correctness levels were high – above 90% – for six of the eight 1-mark items and two of the five 2-mark items.
- Spelling correction improves the performance of the system, though not by much.
- The figures for the 200 sample answers are very similar to those for the full marking trial sample, indicating that generally the 200 answers chosen to help with pattern writing were sufficient to cover the range of answers in the larger sample.

It might be supposed that the system would get the marks right for 100% of the 200 sample answers, since the Oxford researchers wrote the patterns with these answers in mind. However, the patterns are designed to extract key pertinent information from the answers and are not simple copies of the answers themselves; the challenge is to write patterns that are general enough to match paraphrases of the intended content but not so general as to also match unintended content.

Table 1: Average percentage of answers-per-item correctly marked automatically

Item tariff	*Average % correct per item*			
	n	*Corrected*	*Uncorrected*	*200 sample*
1 mark items	3519	92.1	90.7	94.4
1 mark items without 1_7 & 1_8	2744	96.0	94.7	97.9
2 mark items	1793	81.8	79.2	79.3

There is not an obvious reason why the system performed less well on some items than others. It does not seem to depend simply on the facility (difficulty) of the item or the number of alternatives given in the examiners' written marking scheme – item 1_6 (94.5% correct) contained more alternatives than item 1_7 (77.3% correct), whereas item 1_1 (97.5% correct) contained fewer alternatives than 1_7. However, it would seem reasonable to suppose from the results given in Figure 1 that if marking accuracy approaches 100% for the 200 sample answers, the accuracy will also be high for the larger sample of answers. In this way, the sample answers used for writing the patterns might be used to screen for items likely to be unsuitable for automatic marking.

Tables 2 and 3 indicate the correctness of the automatic marks for the spelling-corrected answers, broken down by the correct mark. The percentages relate to the column totals, i.e. in Table 2, 94.1% of the 784 answers with a correct mark of 0 were correctly given a mark of 0 by the automatic system. Key points to note are that:

- The automatic marker is a little more likely to under-mark (give too low a mark) than over-mark. This implies that for these data the system is more likely to fail to credit an unusual right answer than it is to be 'fooled' by a wrong answer.
- For the 2-mark items, very few responses received an automatic mark that was wrong by the full two marks.

2. Strictly, this is the percentage of the unseen answers for which we have undisputed human marks that were marked correctly by the system. See footnote 3.
3. Two approaches were used for determining the 'correct' marks. For the 200 sample answers used by Oxford, the Oxford researchers made their own judgement of what the correct mark should be. For the unseen answers, we considered that when all three human marks agreed, the mark was definitive; we therefore only used these definitively marked answers for the analyses involving 'correct' marks.

Table 2: % correct by mark – (1 mark items)

		Correct mark	
		0	*1*
Auto	0	**94.1**	7.9
	1	5.9	**92.1**
	All	100.0	100.0
	n	784	2735

Table 3: % correct by mark – (2 mark items)

		Correct mark		
		0	*1*	*2*
Auto	0	**89.5**	17.7	2.3
	1	10.4	**77.0**	18.0
	2	0.1	5.3	**79.7**
	All	100.0	100.0	100.0
	n	733	666	394

Inter-marker agreement: Oxford patterns

No matter how careful or well trained they are, human markers inevitably make occasional mistakes[4]. Moreover, some candidates write ambiguous answers that leave room for markers to differ in their expert opinion of whether an answer fulfils the requirements for a particular mark to be awarded – indeed unbiased marking requires that for an answer on the cusp of being worth a mark, half of a hypothetical population of all markers would award the mark and half would withhold it. Of the 6,357 unseen answers, at least two human markers differed in the marks they awarded to 1,047 of them. Table 4 gives pairwise agreement percentages between every pair of markers. For example, the top left cell shows that on average the live and automatic marks exactly agreed for 87.6% of the answers to one-mark items.

4. Marking reliability is controlled through quality control checks – both during marking and also prior to results issue, when checks may be targeted on candidates close to grade boundaries – and by virtue of the fact that many mark-decisions contribute to a candidate's final result.

Key points to note are:

- 1-mark items:
  - When the two items previously identified as outliers (1_7 and 1_8) are ignored, the average exact agreement rates for the auto-human pairs are broadly similar, though a little lower, than for the human-human pairs;
  - All the figures are higher when the outlier items are removed, implying that the human markers were also more likely to disagree on these items – though the differences are smaller for the all-human comparisons than for the auto-human comparisons.
- For the 2-mark items, the average exact agreement rates for the auto-human pairs were lower than for the human-human pairs, though none fall below 70%.The averages reported in Table 4 mask considerable variation by item, however. For example, the auto-marker agrees more often with the live mark for item 2_5 (91.8%) than do either of the human markers (85.1% and 89.7% respectively). There is no obvious reason for this.

Cambridge Assessment's patterns

Patterns for three items were written by the temporary worker recruited by Cambridge Assessment. The accuracy of the automatic marker when using these patterns, compared with the Oxford patterns, may be found in Table 5. Table 6 gives the inter-marker agreement figures. The results for the two sets of patterns are very similar.

Table 5: Percentage of answers correctly marked automatically using Oxford & Cambridge Assessment patterns

Item	*% correct (uncorrected spellings)*		
	Oxford	*Cambridge Assessment*	*n*
1_3	92.4	94.2	449
1_4	97.2	95.3	465
2_3	91.2	93.4	385

Conclusion

We evaluated the automatic marker using eight 1-mark items and five 2-mark items. The items were all taken from a GCSE Biology question paper, and answers from a sample of paper scripts were keyed into a computer file for automatic marking.

The automatic marker marked all but two of the 1-mark items with a high degree of correctness – more than 90% of the answers for which we had a definitive (undisputed) human mark were marked correctly.

Table 4: Pairwise agreement between markers

Item tariff	*Average % exact agreement per item*					
	Live v auto	*Live v exr1*	*Live v exr2*	*Auto v exr1*	*Auto v exr2*	*Ext1 v exr2*
1 mark items	87.6	89.5	93.5	87.6	87.9	92.7
1 mark items without 1_7 and 1_8	91.6	91.8	95.1	92.3	92.5	94.8
2-mark items	72.7	80.1	81.7	73.7	76.6	87.1

Table 6: Inter-marker agreement, Oxford and Cambridge Assessment patterns

Item	*% exact agreement (uncorrected spellings)*					
	Oxford			*Cambridge Assessment*		
	Live v auto	*Auto v exr1*	*Auto v exr2*	*Live v auto*	*Auto v exr1*	*Ext1 v exr2*
q2biii	91.1%	91.1%	89.4%	93.2%	92.8%	91.1%
q4a_fur	91.3%	93.6%	90.6%	89.2%	92.3%	88.4%
q2cii	71.2%	82.9%	85.9%	72.4%	88.2%	87.9%

Agreement levels between the automatic marker and human markers were also broadly similar – for these items – to those found between human markers. We could find no simple explanation for why the remaining two 1-mark items were marked less well by the system – suitability for automatic marking does not appear to depend simply on item difficulty or the number of alternatives given in the examiners' written marking scheme. However, the 200 sample answers used for pattern-writing appear likely to be sufficient for screening 1-mark items for automatic marking. The system was generally less often correct, and there were bigger differences between auto-human and human-human agreement levels, for 2-mark items.

Patterns were written for three of the items by a temporary worker recruited by Cambridge Assessment. This worker was highly qualified in psychology and computing, but had had no previous exposure to the project or computational linguistics. The correctness and inter-marker agreement levels were similar for both sets of patterns, implying that it is possible to transfer pattern-writing skills from the developers to new staff. This is an important step for the commercialisation of the system.

We conclude that automatic marking is promising for 1-mark items requiring a short, textual response. More work is needed to see how the findings generalise to subjects and qualifications other than GCSE Biology, and to investigate why some items are less suitable for automatic marking using this system than others.

References

Pulman, S. G. and Sukkarieh, J. Z. (2005). Automatic short answer marking. In *Association for Computational Linguistics: Proceedings of second workshop building educational applications using NLP*, Ann Arbor, Michigan, 9–16.

Sukkarieh, J. Z., Pulman, S. P. and Raikes, N. (2005). Automatic marking of short, free text responses. *Research Matters: A Cambridge Assessment Publication*, **1**, 19–22. http://www.cambridgeassessment.org.uk/research/

EXAMINATIONS RESEARCH

The curious case of the disappearing mathematicians

John F. Bell and Joanne Emery Research Division

It is not unusual for claims to be made that some aspect of education is getting worse. Mathematics is often cited as a particular area of concern. There have been a number of reports about this issue including Roberts (2002), Smith (2004) and the UK Mathematics Foundation (2005). The declining number of A-level mathematicians is often cited as a particular concern, for example, in the *Times Educational Supplement* Gardiner (2006) wrote

> *'the number of A-level Mathematics students has slumped from 85,000 in 1989 to 66,000 in 2001, and (thanks to the misconceived Curriculum 2000 reforms) to just 52,000 in 2004.'*

A simple calculation would suggest that there has been a fall in numbers of the order of 33,000 students taking A-level mathematics, that is, a 39% decline. However, the interpretation of educational statistics is not a predictable 'one-piece jigsaw' but is instead a fairly simple multi-step problem. The first step is to identify the source of the statistics and check that they are comparable. It is not surprising or unreasonable that the source is not given in a newspaper story. However, an inspection of the available statistics would suggest that no identical definition of A-level mathematics students could simultaneously give a number as high as 85,000 in 1989 and as low as 52,000 in 2004. To investigate this problem, we decided to use the Summer Inter-board Statistics which have been compiled for A-level since 1990 in their present form (some earlier figures were obtained for 1989 but these may be a slight undercount).

After identifying a comparable source of statistics, the next issue is to consider the definition of A-level mathematics students. It is reasonable to assume that from the point of view of Higher Education and employment this should be based on the number with passing grades (A-E). This is important because in 1989 30% failed A-level mathematics and this was only 4% in 2004. A change in failure rates is unsurprising given that the introduction of modular A-levels led to candidates dropping mathematics rather than completing the course, obtaining a U and appearing in the statistics. Another relevant factor is the number of 17-year-olds in the population. This varied considerably over the period in

Table 1: Statistics for successful A-level mathematicians

Year	*A-level Maths passes*[1]	*Further Maths passes*	*No. of 17 year olds (government estimates in thousands)*[2]	*Maths passes/ No. of 17 year olds × 100*	*Further Maths passes/ No. of 17 year olds × 100*
1989[3]	50,570	5,022	715.2	7.07	0.70
1990	53,954	5,314	644.4	7.91	0.78
1991	51,185	5,144	599.3	7.94	0.80
1992	50,530	4,826	576.0	8.43	0.81
1993	50,129	4,553	551.5	8.70	0.79
1994	50,988	4,271	529.4	9.25	0.77
1995	51,708	4,465	537.2	9.77	0.84
1996	54,674	5,086	582.7	10.18	0.95
1997	58,514	5,216	604.0	9.68	0.86
1998	56,270	5,540	600.8	9.32	0.92
1999	56,192	5,306	586.2	9.35	0.88
2000	54,243	5,164	593.7	9.25	0.88
2001	54,193	5,223	614.7	9.13	0.88
2002	45,398	4,819	643.1	7.39	0.78
2003	47,059	5,224	645.4	7.32	0.81
2004	49,052	5,620	644.4	7.60	0.87

1. All mathematics except further mathematics 2. Population estimates – current releases – datasets T-04: England 3. These figures may be a slight undercount

question. In 1989 the cohort size was 715,200 and in the 2004 it was 645,400 and had fluctuated in between. To control for this two indices were calculated: the ratio of A-level mathematics passes over the number of 17-year-olds times one hundred and the same formula for further mathematics. Note these are not percentages since not all A-level candidates are in the 17-year-old age cohort.

The data derived from the above sources and calculations are presented in Table 1. The second column shows the number of A-level passes for any mathematics except further mathematics. This has not declined massively but has been around 50,000 for the period covered. Furthermore, it increased in the late 1990s and declined with the first set of A-level results in 2002 arising from the introduction of Curriculum 2000. The next column is the number of successful candidates for further mathematics. These have varied by around 5,000 and have shown no sign of a decline. In the fourth column, estimates of the number of 17-year-olds have been presented. The next two columns are the indices that control for cohort size. The index for mathematics passes increased until 1996 then started to decline. However, the latest available figure is not that much different from the figure in 1990. There is no evidence of a spiral of decline or an impending catastrophe. For further mathematics, the trend is less clear but there are now more further mathematicians than there were in 1989.

It is clear that there was a decline in mathematics passes associated with the introduction of Curriculum 2000. This curriculum change meant that typically candidates started studying four or five A-levels and then at the start of the upper sixth chose three of them to study for A-level. This has created problems for the interpretation of statistics about examinations. The awarding of modular A-levels is based on the results for candidates who have taken sufficient modules to obtain AS and A-levels. However, candidates do not necessarily claim the qualification preferring to resit modules or not bother if the overall grade is a U or they are taking the subject as an A-level. Thus, it is possible to have candidates in the matched databases who have AS only, both AS and A-level, and A-level only (although these candidates could have claimed an AS).

To understand the process it is necessary to analyse large matched databases of individual examination results (these are generated for England only to provide data for performance tables). By analysing these databases, we calculated that in 2003 there were approximately 15,000 year 12 candidates in England who obtained a passing grade at AS but did not proceed to A-level (note the numbers in Table 1 contain results for all ages and Wales and Northern Ireland as well as England). The total with a combination of AS and A levels was approximately 53,000. In 2000 there were approximately 4,000 candidates who succeeded in AS only (strictly, this is a different qualification, the Advanced Supplementary rather than the Advanced Subsidiary) and only 47,000 with either combinations of AS and A-levels. The introduction of Curriculum 2002 has been a success in increasing the numbers studying mathematics beyond GCSE. The decline in A-level success as a result of Curriculum 2000 can be explained by candidates not opting to take the A2 modules. One possible explanation is that there were several thousand candidates in the years prior to 2002 who would have dropped mathematics if they had been given the option and that these candidates would not have opted for a highly numerate discipline in higher education.

Another issue that is often raised is how to increase the number of A-level mathematicians. QCA (2006) argues that it is necessary to make A-level mathematics appeal beyond a 'clever core' of mathematicians to less enthusiastic and less able students. However, Gardiner (2006) criticises this and argues that there is a pool of able students:

> *'There are 31,500 students achieving A* grades in GCSE maths, yet the authors [referring to QCA, 2006] have no idea how many of these take maths A-level. One might expect between 10,000 and 15,000 to go on to A-level (the current number is probably much lower), and one can imagine incentives that would increase this to 20,000-plus.'*

Gardiner's estimates convert to between 31% and 48% for his expectations and 63% for his target with incentives. It is possible to calculate the figure using the matched database for England only. For those taking A-levels in 2004, the percentage of those obtaining A* going on to take A-level mathematics is 62%. To put this into perspective, this is higher than equivalent percentages for other subjects, for example, Chemistry (51%), English (48%), Biology (43%), Geography (38%) and

French (37%). In addition, 85% of successful mathematics students who progressed from GCSE mathematics had an A or A* at GCSE mathematics. This is comparable with some other subjects (for example, French – 93%, the sciences - ~80%) but is much higher than for other subjects such as English – 50% and history – 49%. These figures support the argument that if there is to be a large increase it must come from beyond the 'clever core'.

The findings of this research may come as a surprise. There has been no large scale decline in the number succeeding in A-level mathematics. The disappearing mathematicians can be accounted for by changes in the structure of mathematics leading to candidates dropping out rather than failing, and to demographic changes.

References

Gardiner, T. (2006). Make maths count. *Times Educational Supplement*, 10 March 2006, p.21. http://www.tes.co.uk/search/story/?story_id=2206638

QCA Research Faculty (2006). *Evaluation of the participation in A-level mathematics interim report. Autumn 2005*. London: QCA. http://www.qca.org.uk/downloads/qca-06-2326-gce-maths-participation.pdf

Roberts, G. (2002). SET for success. *The supply of people with science, technology, engineering and mathematics skills. The report of Sir Gareth Roberts' Review*. London: HM Treasury. http://www.hm-treasury.gov.uk/documents/enterprise_and_productivity/research_and_enterprise/ent_res_roberts.cfm

Smith, A. (2004). *Making mathematics count. The report of Professor Adrian Smith's inquiry into post-14 mathematics education*. London: DFES. http://www.mathsinquiry.org.uk/

UK Mathematics Foundation (2005). *Where will the next generation of mathematicians come from?* Preliminary Report of a meeting held in Manchester. 18/19 March 2005. http://www.ma.umist.ac.uk/avb/pdf/WhereFromPreliminaryReport.pdf

STANDARDS OVER TIME

What happens when four *Financial Times*' journalists go under the eye of the invigilator?

Miranda Green *Financial Times*, Education Correspondent, additional research by Anna Metcalf

In the weeks leading up to A-level results day Cambridge Assessment Research Division and its UK examination board, OCR, worked with *The Financial Times* to illustrate certain aspects of A-levels. On Saturday August 20, 2005, *The Financial Times* published the following two articles in *FT Weekend* which are reproduced here with their permission.

Four *FT* experts, and four, surely, of the most awkward and disputatious candidates ever likely to grace an examination hall. Earlier this summer, as the exam results season approached and, with it, the inevitable annual debate over standards, *FT Weekend* had a bright and apparently simple idea for getting to the truth behind the increasingly ritualistic argument: why not get a handful of the *FT*'s brightest and best to sit some of this year's papers? These writers, who live, breathe and even, dare we say it, pontificate about the subjects under consideration every day of their working lives, would then be able to give us their impressions of how today's exams compared with those that crowned their own school years.

Several twisted arms later, Lucy Kellaway, work columnist, James Blitz, political editor, Chris Giles, economics editor, and John Lloyd, editor of the *FT Magazine* and commentator on the media, had agreed to face their demons, and possible public ridicule, by submitting to an ordeal that most of the experts, politicians and critics who annually bemoan falling standards have long put behind them.

Accusations of dumbed down questions, grade inflation and lenient marking have dogged the A-level, once the unassailable 'gold standard', for years now. Every August, opposition MPs, employers, universities and independent schools voice their suspicions that ever-higher results (2005 is the 23rd year of improved pass rates) do not represent a true step forward in education or attainment. Their comments are reported just as families are waiting anxiously for an envelope from the exam board. Teachers and government ministers then reproach the doom-mongers for casting a cloud over the latest crop of good results, which they insist have been fairly earned by hard-working students.

But this year the pass rate edged up again to 96.2% – with A grades up to 22.8% – and the exam watchdog, the Qualifications and Curriculum Authority (QCA), has admitted that it will hit 100% in the near future.

If all candidates sitting an A-level – still well below half of the age-group – are deemed worthy of an E or above, that will deal a fatal blow to the credibility of certificates, say critics of the system. But, at bottom, the debate is about what an A-level with a near-universal pass rate is measuring and how marks and grades – particularly the A on which top universities have traditionally relied to 'sort' applicants – are awarded. It was this issue that our 'volunteers' agreed to probe.

"They're not going to ask me about Plato's *Republic*, are they?" said an anxious James as he agreed to put his strengths as the leader of the *FT*'s Westminster team to the test by sitting a paper on UK government.

Chris – economics editor of "the world's business newspaper" and preparing to answer questions on the national and international economy designed for 17 and 18 year olds – spotted the downsides with the ruthless insight for which he is feared and famed: "O God, go on then. Put me down for humiliation."

Lucy and John, both parents of exam-age children, took more of a scientific interest and approached their papers – on business studies and media studies respectively – with rather more equanimity.

After much debate about how to carry out the experiment, we decided to work in co-operation with an exam board, largely because we would

then be able to see the markers' comments, giving us more insight into our guinea pigs' performance. Cambridge Assessment – formerly the Oxford and Cambridge board – was quick to caution that if we were expecting our experts to romp home to good grades without any preparation, thus 'proving' that standards had slipped, we were likely to be disappointed. Mindful of Margot Fonteyn's comment at the height of her career that she would probably fail elementary ballet if she had to retake it, we had no such preconceptions.

"All the research demonstrates that older people taking exams targeted at a particular younger age-group tend to do badly – because they simply know too much," said a gleeful Bene't Steinberg, the public affairs director charged with defending the board's standards against the annual accusations of dumbing down.

His colleague Sylvia Green, head of research, explained in more detail: "The journalists will have the advantage of their life experiences but will also have the disadvantage of the passage of time since they studied."

By Sylvia's reckoning Chris, at 35 our youngest candidate, had the edge simply because his was the most recent experience of exam conditions, and there had been less time to spoil his chances of jumping through academic hoops.

"Students do not have the knowledge base of journalists and this may well lead to greater focus on the course curriculum and less writing 'off-topic'," she said, warning our bright-eyed volunteers not to get too clever.

"The journalists' broader knowledge may also lead them to have more sophisticated expectations of the questions and perhaps to misjudge their complexity."

Later, we discovered that internal betting among the examiners predicted that it would be impossible for any of our writers to get above a C grade. (We kindly refrained from imparting this information to our candidates until after the test.)

Core subjects such as maths, English, science or languages were ruled out because previous experiments have shown only that someone – pretty much anyone – coming cold to an exam based on a taught curriculum and syllabus that can change every year will struggle and do badly.

So we decided to focus on the more analytical, theoretical subjects, which would have the advantage of testing a more interesting and thorny question and one much debated by universities, employers and head teachers: do A-levels still reward critical thought and well-expressed argument as well as the accumulation of information?

To make our guinea pigs feel comfortable, their invigilator was no one more intimidating than myself, the education correspondent; the exam hall was a meeting room at HQ; and organic apples were laid on to help with dipping blood sugar. A somewhat dusty energy bar was also claimed from the bottom of my handbag by James, and John asked for a double espresso during the second hour of his ordeal (he was sitting the longest paper).

As nervous candidates do, they compared notes on how they had prepared. Most of our writers had done little work beforehand, but Lucy had been given sight of an extract for the business case study on which she was to be examined: the fascinating tale of McAvoys Ltd, a Scottish fish farm, and its troubled quest for growth. Experts have suggested that this practice of giving students an early look at some of the exam material does theoretically reduce the 'demand' or difficulty of questions, but in previous years it was felt that setting unseen text passages resulted in too many rewards for comprehension and not enough for understanding of the subject. As with many of the technical conundrums surrounding A-level standards, there are arguments on both sides.

Chris admitted to having looked at some of the specimen papers and the mark scheme on the website, which promptly led to accusations from the others of being a swot.

But the economics editor hadn't been impressed. "I was rather shocked by these as they contained clear errors," Chris said. "One set of answers had highly dubious suggestions such as 'an economic slowdown in Japan would lead to a reduction in the output and employment of Japanese companies based in the UK.'"

The hour had come and I switched into invigilator mode. "Read the question, read the rubric," I reminded them sternly as the papers were turned over.

This was the moment when the experience became 'real' and, judging from the grunts, sighs, head scratching and pained faces, it was challenging. At one point Lucy's pallor and James' despairing look at the sky made me fear they were about to duck out, but they stayed the course.

Afterwards, I asked for their impressions. Chris was disappointed by the extent to which questions asked for simple facts and definitions and by the easy calculations required. "I was surprised and concerned by the limit of historical knowledge that was expected of students (only 10 years of economic history) and the extent to which the papers encouraged parroting of facts rather than eliciting whether students had understood the relevant issues."

The business studies paper prompted the same sort of objections from Lucy, who thought factual questions on regional aid and company finance "boring" and too clearly designed to elicit a particular answer: "I found myself searching my brain for any scraps of knowledge that might do. I find it very hard to assess if these were by some miracle correct, or if they were looking for something else entirely."

Here she hit a nerve. Official analysis carried out for the QCA of the new-look A-levels introduced in 2000, when the traditional two-year courses were split into AS and A2, also identified this problem. The level of detail required by the examiners, and the very explicit guidelines, supporting materials and published marking schemes might lead to a monotonous time for pupils and teachers as they churned out what was expected of them, the reviewers found.

Schools might "concentrate on delivery of the specification, relating it heavily to the anticipated assessment and giving candidates a narrower learning experience".

This warning, together with Lucy's argumentative assertion that if she took the paper again she would know from the markers' comments how to get herself an A, seems to confirm that the approach of a 100% pass rate might be in part due to what is known as 'teaching to the test'.

The QCA disputes that this is a widespread problem, unlike in school tests for 7, 11 and 14 year olds. But head teachers freely admit that their staff have become very savvy about the tricks which will secure top grades for pupils.

In contrast to Chris and Lucy's complaints about boring questions and – once they had received the results and seen the comments – "pernickety" marking, John and James were more impressed by the demands the papers made. John found himself pleasantly surprised by how much his essay questions, on issues such as media regulation and competition and the British film industry, required in terms of thought, by their relevance to current debate and by the experience itself. He was, however, comparing this year's A-level paper to his own Scottish

Highers, which are a slightly lower standard because more subjects are taken.

"I know it sounds a bit sick but I quite enjoyed it," he said, handing in the final script. "I didn't find today's questions very difficult but my fear is that I waffled and that there is still such a thing as the 'right' answer, which I didn't know."

Matthew Lumby of the QCA confirmed that this fear was rational, because essays are marked in a very structured way: "A lot of people think that in an essay question you are just judged on content and style when in fact the markers will be looking for a number of specific things."

James was my biggest worry. He arrived late, he was distracted by a big political story he had to write that day, he finished early and he seemed, bizarrely, to be writing in pens of about three different colours. The questions would have been acceptable for any political correspondent, he decided, and – falling into the trap identified by Sylvia of underrating the sophistication of the paper, as we later discovered when the results came in – he declared that even a 17-year-old could cope merely by reading the newspapers closely. "But this experiment, never to be repeated, has made me think that A-levels aren't a cinch," he added. "That was a lot to do in one hour."

As James and Lucy seemed to feel acutely, the experience revived horrible memories of both school and university and, employers' organisations will be interested to note, was utterly different from working life – irrelevant really.

"An exam is a limited and stressful way of testing people's ability. You can learn to be better at exams but why bother?" said Lucy afterwards, with renewed empathy for the pupils who are yearly told their certificates are all-but-worthless after being put through this "wretched" experience.

"I doubt I've sat down and written so much with a biro since university – that was the really hard bit," said James. (As it turned out, he was right to be concerned about being penalised for handwriting violations.)

According to our exam board experts, having nothing riding on the exam and not taking it very seriously was likely to have had a negative impact on his performance as well: empirical studies have shown that morale, defined as optimism and belief in the exam's link to a successful future, is crucial.

James preferred to blame the others for his lack of concentration: "I kept getting psyched out by Lloyd's seriousness but that's nothing new. I thought Lucy ate her apple a bit noisily."

John, meanwhile, was preparing his own self-justifications as the candidates compared grade predictions.

"I think if there is justice in the world I should get an A star. Lucy said that's what men always do, overestimate their prowess while women are more modest. I claim that's the false modesty of one who knows she's done well. That's what the girls in my school were like."

As far as predictions went, both were right: all three male candidates expected to be awarded an A, but only two achieved the highest grade (see below). Lucy expected "no higher than a C" but came away with a B and is now claiming she could have done better if the marker had read her answers more carefully.

But overall, exam board experts and educationalists have confirmed, our candidates turned in a very creditable performance. John is even talking about taking Russian or Italian A-level next, and all felt they learnt something from the experience – if only never to answer my calls in future.

A couple of hours after the exams, Lucy complained in a one-line e-mail that it would take years to repair the damage the experience had done to her self-esteem and mental health.

It is likely to take at least as long for the government to decide whether it wants to keep the A-levels it still thinks are the most recognisable qualification in the education system or back proposals for a complete overhaul.

In the meantime, the debate will continue to rage, year in, year out, with the two sides taking increasingly entrenched positions based largely on opinion and prejudice. For the truth is that it is extremely difficult to compare today's exams with those from more than two decades ago, when only a limited quota of students could be awarded an A. An independent study of standards over time, carried out by an international panel, was flummoxed when faced with the numerous changes to the system and to the tests of skills and knowledge since the 1980s.

But our experiment seems to suggest that a pupil who has been spoonfed facts and definitions, and reproduces them in exactly the phrasing the examiner expects, will indeed find it easier nowadays to do well on the structured parts of the papers: our candidates were right to fear that without being trained in the curriculum content they would not deliver the right answer in the right way.

However, the sort of skills that both universities and employers say they look for – the ability to think quickly and critically on your feet and make a clear argument – are also clearly still being sought and rewarded. Our most successful *FT* candidates felt they were more generously marked on the essay questions, which, because of their profession, they were able to answer fluently and cogently.

In fact, the markers who awarded the two lower grades complained that it was exactly these elements that were missing. Answers were not "wide-ranging" enough or did not provide enough analysis to support a judgment, they noted.

Rising to the challenge offered by essay questions and using them to show off both skills and knowledge still, it would seem, secures the best marks.

As Chris, the best prepared of our candidates, reflected: "I didn't think my 1988 A-level was perfect, and nor was this. The multiple choice in 1988 was harder than the structured questions in this paper and you had to write more essays. Those aspects were certainly easier now. On the other hand, there was no choice of questions on this paper so students were no longer able to pick their favourite subject to the same degree as we could in the 1980s."

‘I expect to get an A’: How the *FT* writers thought they would do – how they actually did.

James Blitz, Chris Giles, Lucy Kellaway and John Lloyd *Financial Times*

Candidate name: James Blitz

CANDIDATE OCCUPATION: political editor, *FT*

PAPER: Government of the UK (AS-level Politics)

CANDIDATE INPUT: no preparation; one apple, one Food Doctor energy bar

OUTPUT: 11 pages of barely legible, multicoloured scrawl

PREDICTION: “I expect to get an A. I think the questions were not difficult for a political correspondent, and dare I say they were a cinch for someone who did the three-hour Oxford M. Phil paper on Marxist Leninist social and economic theory. “

RESULT: Grade C (63 out of 100)

MARKER’S COMMENTS: “The candidate clearly has a good understanding of contemporary government and politics but lacks the detailed knowledge and technique required to access the higher levels of the assessment matrix at A-level. The legibility of the answers is an issue: marks are awarded for communication and significant parts of the script, particularly towards the end, are indecipherable.”

Candidate name: Christopher Giles

CANDIDATE OCCUPATION: economics editor, *FT*

PAPER: The National and International Economy (AS-level Economics)

CANDIDATE INPUT: previous papers, specimen answers and mark scheme researched; one apple and one tuna sandwich consumed

OUTPUT: six pages of meticulously completed answers with minimal crossing out

PREDICTION: “I will not be able to hold my head up high if I do not get an A. I mean if I do not know ‘the economic consequences of inflation’ – my essay question – I should not be in the job I am.”

RESULT: Grade A (39 out of 45)

MARKER’S COMMENTS: “A strong performance. The candidate answered questions directly, paid attention to the directive words and applied relevant macroeconomic terms and concepts. On a few of the questions, however, he could have explained the points he made rather more fully and analytically.”

Candidate name: Lucy Kellaway

CANDIDATE OCCUPATION: work columnist, *FT,* and author of *Who Moved My BlackBerry?*

PAPER: Businesses, Their Objectives and Environment (AS-level Business Studies)

CANDIDATE INPUT: perusal of a one-page brief on a business case study; one apple and one wholenut chocolate bar consumed

OUTPUT: six pages of neat script and concise answers, one question attempted and then crossed out

PREDICTION: “I think I’ve got a C. If I get less than that I’ll be ashamed and outraged, as in the end I did manage to think of something to say and didn’t leave anything out.”

RESULT: Grade B (31 out of 45)

MARKER’S COMMENTS: “The candidate demonstrated a sound understanding of business issues on this paper in almost every question but would have scored more highly if the ‘trigger’ words in the question had been used as a guide as to how to approach the answer.”

Candidate name: John Lloyd

CANDIDATE OCCUPATION: editor, *FT Magazine*, and author of *What the Media are Doing to our Politics*.

PAPER: Media Issues and Debates (A-level Media Studies)

CANDIDATE INPUT: Sample exam paper read, children were asked for help but this was not provided; one apple and one egg sandwich consumed

OUTPUT: $13^{1}/_{2}$ densely covered pages of essay questions

PREDICTION: “If there is any justice in the world I should get an A star. I was interested at how connected to current issues the media studies paper was – and how much it demanded in terms of thought.”

RESULT: Grade A (60 out of 60)

MARKER’S COMMENTS: “There could be an argument for deducting marks on the basis of omissions. But to do so would be churlish in view of the quality of the answers in other respects. Faced with these answers in the exam season, the candidate would receive maximum marks in each case.”

The Cambridge Assessment Network

Andrew Watts Cambridge Assessment Network

The mission of the Cambridge Assessment Network is to become 'a virtual centre of excellence' for professional development in assessment. It was launched in October 2005 at the first, very successful Cambridge Assessment Conference. The theme of the conference was 'Maintaining Trust in Public Assessment Systems', which indicated the kind of underlying issue in assessment that the Cambridge Assessment Network aims to address. The Network recently launched its internal, virtual learning environment, Campus, for Cambridge Assessment staff. This is an informal communication channel through which staff can share knowledge via discussion boards, chat rooms and other online activities. At the beginning of April a similar external website, AssessNet, was launched for those assessment professionals, both in the UK and internationally, who want to belong to a world-wide assessment network.

The Research Division and the Cambridge Assessment Network have already established a close working relationship. The work of researchers is a prime source of new material for Network seminars and workshops, and the Network supports the dissemination of research ideas throughout the Cambridge Assessment Group. In Autumn 2005 a short series of seminars was run in which all members of the Research Division participated. These were well attended in both Coventry and in Cambridge, where we had to move out of our own buildings to bigger rooms because of the numbers wishing to attend. Colleagues from all business streams appreciated the opportunity to keep up-to-date with current thinking and to think through ideas beyond their immediate sphere of work. A format of having three short presentations in one longer seminar, with discussion between each and at the end, proved informative and stimulating.

In meeting up in such events with colleagues from across the Cambridge Assessment Group, members of the Research Division benefit from the comments and suggestions of those who are engaged in the day-to-day work of developing assessments. Thus such seminars can be a forum in which material is introduced which is being worked up into a publication or conference paper. The interaction with colleagues can be a first outing for ideas which will be refined later in the light of the discussion.

Members of the Research Division have a particular part to play in the Cambridge Assessment Group, since the opportunities they have had to work in different parts of the business often give them a broad view of what is happening. They can thus describe for us the bigger picture and also remind us of the underlying principles on which our work is based. An example is Tom Bramley's seminar series on 'The Basics of Assessment'. This has been run in both Cambridge and Coventry/ Birmingham and has proved very successful, with average attendance of over 25 at the sessions. Many of those who attended commented on the usefulness of being able to stand back and think about how fundamental concepts apply to their work. Another example of research work put to Group-wide use is the Question Writers' Interactive Learning Tool (QWILT), which continues to be used from time to time in the training of examiners by colleagues in OCR, both general assessment and vocational, and in CIE. Last summer Victoria Crisp, from the Research Division, and Andrew Watts ran a one-week workshop for officers in the Singapore Ministry of Education using that material.

On a broader front, members of the Research Division have led the way in promoting a culture in which it is expected that we share what we know and that from time to time we will present aspects of our work to our peers. This kind of activity is fundamental to the successful running of a lively community of practice. The Cambridge Assessment Network aims to encourage such participation as a way to establish a Group-wide learning culture in Cambridge Assessment.

The Cambridge Assessment Network and the Research Division also work together to disseminate new ideas in assessment and to help colleagues to keep in touch with what is developing in the world of assessment. In the Network's 'New Horizons' seminar series innovations are discussed with colleagues from across the Group. One such issue is the use that will be made of item level data from public examinations, once it becomes more available through greater use of on-line collection of data and the electronic management of examination scripts.
The Network ran a seminar on this for members of the Group in April.

When it comes to informal networking, research staff have been able to contribute particularly because they have information that is perhaps not available to others or they have a different perspective. This helps the kind of exchange and the kind of exploration of issues across business streams which is one of the main objectives of the Cambridge Assessment Network. Members of the Division have also been active in chairing discussion groups and seminars, in facilitating cross-business stream discussions, and in introducing visiting speakers.

The Cambridge Assessment Network believes strongly in the value of informal meetings, and those with research colleagues have proved very beneficial. As Network staff have prepared materials and written courses and papers, informal discussions have taken place, materials have been exchanged and ideas refined. The benefits of informal exchanges are not just one-way since Cambridge Assessment Network writers get help with their materials and research staff explore how to communicate their ideas effectively to those working in other areas.

Finally, another opportunity for working together took place at the end of March when the Cambridge Assessment Network introduced a seminar at Transport House at which members of the Research Division presented some significant work on aspects of students' writing in English examinations.

All in all, the links between the Research Division and the Cambridge Assessment Network are already significant. We believe that it is highly beneficial for our Research teams to have different channels of communication to different audiences both in the Group and beyond. Also, it is important that research can feed in to the professional development of those who want to improve their understanding of assessment issues. The Cambridge Assessment Network also benefits, by having close at hand a source of lively ideas and high quality research, as well as a team of colleagues who are willing to present what they know and discuss it at events for their colleagues from the wider Group.

Research News

Conferences and seminars

Cambridge Assessment Conference

The inaugural Cambridge Assessment conference 'A Question of Confidence: Maintaining Trust in National Assessment Systems' took place at Robinson College on 17 October 2005. The main speakers were Baroness Onora O'Neill, Professor Barry McGaw, Dr Nicholas Tate and Professor Alison Wolf. The topical theme of the conference and the high calibre of speakers attracted over 200 delegates from the UK and around the world, representing schools, colleges, universities, awarding bodies and educational organisations.

10th Roundtable Conference, Melbourne, Australia

In October 2005 Sylvia Green was invited to give a keynote address at the 10th Roundtable Conference in Melbourne. This is an annual conference on student assessment and measurement issues and provides an opportunity to enhance the professional knowledge of delegates by sharing information about assessment and reporting practices in Australia as well as internationally. Sylvia's presentation was on: 'Harnessing technology in the brave new world: affordances and challenges in e-assessment'.

6th Annual Association for Educational Assessment-Europe Conference, Dublin

Jackie Greatorex, Irenka Suto and Sylvia Green hosted discussion groups at the AEA-Europe conference on *Assessment and Accountability* in November 2005. Topics included 'Gaining theoretical insights into the A-level and GCSE marking process' and 'International perspectives on assessment in vocational qualifications'.

Transport House seminar

In March a seminar on *Variations in aspects of writing in 16+ examinations between 1980 and 2004* was presented by Cambridge Assessment at Transport House, London. The chair was Barry Sheerman, Chairman of the Select Committee on Education. *Variations in aspects of writing* is an extension to the research published by A.J. Massey & G. L. Elliott in 1996, which explored differences in written English in public examinations set in 1980, 1993 and 1994, by the inclusion of a further sample from an examination set in 2004. A copy of the report by A.J. Massey, G.L. Elliott and N.K. Johnson can be downloaded from: http://www.cambridgeassessment.org.uk/research/aspectsofwriting

Institute for Public Policy Research (IPPR) research forum

In April IPPR, with the support of Cambridge Assessment and Select Education plc, presented the second research forum on 'Curriculum, Assessment and Pedagogy: Beyond the "standards agenda"'. This was the second event of a major project, *A New Agenda for Schools*. The event provided an opportunity to think critically about the current relationship between curriculum, assessment and pedagogy in light of the latest evidence from the UK and abroad. Participants were asked to consider future long-term policy options with a view to creating a system which prioritises the need to reduce the attainment gap, broaden learning and make gains in attainment at Key Stage 1 more sustainable as pupils progress through Key Stages 2 and 3. The speakers included Sylvia Green from the Research Division.

International Association of Educational Assessment Conference (IAEA)

Sylvia Green, Jackie Greatorex and Nicholas Raikes, along with other colleagues from Cambridge Assessment, presented papers at the IAEA conference in Singapore in May. The theme of the conference was 'Assessment in an Era of Rapid Change: Innovation and Best Practices'.

Publication

Reading, Writing, Thinking, published by the International Reading Association in 2005, includes a chapter by Martin Johnson of the Research Division. 'Concepts of difficulty: A child's eye view', reports on a project that gathered qualitative data about children's perceptions of writing test stimuli. The book addresses the issues of the power of writing, the assessment of language and literacy development, adolescent literacy, beginning reading and writing, multiculturalism and language learning, and literacy and critical thinking.

'On-line mathematics assessment: The impact of mode on performance and behaviour' by Martin Johnson and Sylvia Green has been published on line in the *Journal of Technology, Assessment and Learning*. This can be downloaded from http://www.jtla.org

An article by John F. Bell and Eva Malacova, 'Changing boards: Investigating the effects of centres changing specifications for English GCSE', was published in the Spring 2006 issue of *Curriculum Journal*.

'Can a picture ruin a thousand words? The effects of visual resources and layout in exam questions' by Victoria Crisp and Ezekiel Sweiry will be published in *Educational Research* in June 2006.

AssessNet

AssessNet, the Cambridge Assessment Network's online networking tool for worldwide assessment professionals, was launched on 31 March 2006. Log on now at http://assessnet.org.uk for the latest education news and resources, details of the Network's events, and the opportunity to network with colleagues via the discussion forums, chat rooms and other exciting features.

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

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