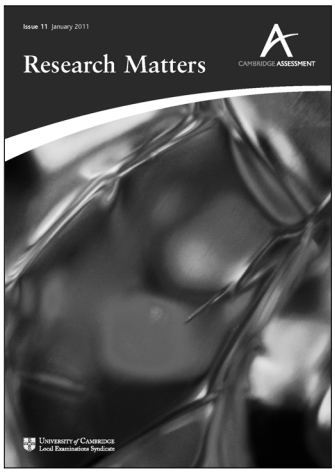
Issue 11 January 2011

Research Matters

CAMBRIDGE ASSESSMENT

University of Cambridge
Local Examinations Syndicate



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

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

Research Matters : 11

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

Research Matters again establishes the importance of a highly elaborated research enterprise around assessment and qualifications. Shaw and Crisp's work on approaches to validity research is consistent with the commitments of the Cambridge Approach regarding the necessity for validation research; Shiell *et al*. drill down into issues of the impact of marking mode on reliability (again a commitment in the Cambridge Approach); and Novaković and Greatorex revisit the issue of demand – this time in the context of vocational qualifications. The work indicates the scale and extent of awarding bodies' responsibility for understanding the operation and impact of the assessments and qualifications which they develop and administer. Such work is not demanded in the quality requirements enshrined in English regulatory arrangements, and shows how meeting strict moral and technical standards extends well beyond the obligations contained in regulation. It is on this that both Vidal Rodeiro and Nádas' work on modularisation and Steinberg and Hyder's work on regulation of standards shed considerable light. The former shows that views regarding 'modular qualifications bad, linear good' – and vice versa – do not reflect the complex ways in which modular and linear qualifications operate in schools and colleges. Policy makers', educators', commentators' and the public's perceptions of the value and application of modular and linear qualifications need to be informed by the realities disclosed by sound research. And hence to regulation – the sum of the analysis presented suggests that the current commitments to regulation which attempts to steer and shape, rather than directly control, all aspects of qualification arrangements will be more possible in a context where awarding bodies undertake comprehensive and incisive research and validation work – leading to a well-balanced set of system arrangements.

Tim Oates *Group Director, Assessment Research and Development*

Editorial

The articles in this issue illustrate the challenges facing the assessment community across a range of areas. In Issue 8 of *Research Matters* Johnson and Nádas reported on their study of on-screen marking and the impact of mode on reliability and marking behaviours. They concluded that further research should be conducted to explore mode-related effects in marking essays of greater length. Shiell *et al*. here report on their latest study where they replicated the research in the context of extended Advanced GCE essays. In their work on the effects of GCSE modularisation Vidal Rodeiro and Nádas combined quantitative and qualitative research methods to investigate the impact of modular assessment on GCSE students, specifically in the key area of *flexible assessment*. This allows units to be taken at the end of a course in a linear fashion or to be taken in different sessions throughout the course to follow a more unitised approach to teaching and learning. Shaw and Crisp then discuss how perceptions of validity have changed over time and the issues that have led to these changes. Their work illustrates the complexity of validity and its importance given the high stakes nature of educational outcomes, their uses and the inferences based upon them.

Black and Gill test the hypothesis that Critical Thinking (CT) skills are transferable and can be applied to other subjects in a beneficial way. They consider some of the research evidence in this field and discuss the best way to deliver CT so as to foster transferrable CT skills and dispositions. Novaković and Greatorex focus on a review of literature, theory and method in their article on the context of vocational qualifications. They consider how instruments used in the vocational field could be used to compare different types of qualifications and the effectiveness of existing methodologies. This is a complex area, fuelled by expectations that standards should remain constant over time, across subjects, between awarding bodies and between task and test demands.

Bramley and Oates describe two research methods that are used within Cambridge Assessment both for operational and experimental purposes. Rank ordering and paired comparison methodologies have been used extensively in the comparability research work of Cambridge Assessment and their use in operational aspects of examinations is being explored and validated. Steinberg and Hyder discuss the need for minimal and useful regulation and how new patterns of engagement between those concerned with the creation and use of assessments can lead to the better regulation of public examinations.

Sylvia Green *Director of Research*

Extended essay marking on screen: Does marking mode influence marking outcomes and processes?

Hannah Shiell, Martin Johnson, Rebecca Hopkin, Rita Nádas and John Bell Research Division

Introduction

Technological developments are impacting upon UK assessment practices in many ways. For awarding bodies, a key example of such impact is the ongoing shift towards examiners marking digitally scanned copies of examination scripts on screen rather than the original paper documents. This digital shift affords opportunities to manage and distribute information in ways that are not possible in paper-based marking systems, and this has important quality assurance benefits.

At the same time, however, the shift towards marking scripts on screen has prompted questions about whether the mode of marking might influence the outcomes of the marking process, particularly in relation to essay responses.

Research into comparisons between how people read texts on paper and computer screen suggests that the medium in which a text is read might influence the way that a reader comprehends that text. This is because some of the reading behaviours that support comprehension building, such as seamless navigation and annotation of text, are not easily replicated on screen (Dillon, 1994; Marshall and Bly, 2005; O'Hara and Sellen, 1997; Piolat, Roussey and Thunin, 1997; Rose, 2010).

Additional research also suggests that reading long texts can be more cognitively demanding on screen (Wästlund, Reinikka, Norlander and Archer, 2005), and that this extra demand can have a detrimental effect on how readers comprehend longer texts (Just and Carpenter,1987; Mayes, Sims and Koonce, 2001). In the context of examination marking, there might be concerns that such a mode-related effect might lead to essays being marked less accurately when marked on screen compared with when they are marked on paper.

The theoretical basis for concerns about mode-related influences on essay marking can be summarised by the model presented in Figure 1. This model outlines the potential relationships that are involved when an examiner reads an essay in order to mark it. In summary, literature underpinning the model infers that the shift from marking essays on paper to marking them on screen might be expected to impact upon examiners' manual and cognitive marking processes. This could, in turn, result in examiners having a weaker comprehension of essays when marking them on screen and this might be reflected in the final marking outcome.

Research in this area is therefore a principal concern for awarding bodies and stakeholders, posing potential implications in terms of both the defensibility of assessment outcomes and public trust in the assessment system.

In response to these concerns, researchers at Cambridge Assessment and elsewhere have been investigating how transition from paper-based to screen-based essay marking might influence examiners' marking behaviours and their marking accuracy. Four recent studies have investigated how mode might affect essay marking (Johnson and Nádas, 2009; Coniam, 2009; Fowles, 2008; Shaw and Imam, 2008). These studies,

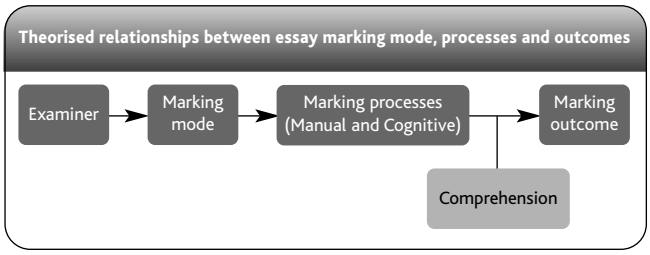


Figure 1 : Theorised relationships between essay marking mode, processes and outcomes

which consider essays of 150 to 600 words, report a negligible mode-related effect on marking accuracy; suggesting little cause for concern as the marking of digital essay images on screen replaces the marking of hard-copy paper essays.

Among the four studies, Johnson and Nádas (2009) is noteworthy as it employs a wider variety of quantitative and qualitative methods. The aim of the project was to broaden investigation beyond the singular consideration of the effects of mode on marking accuracy; to also explore mode-related influences on recognition of essay quality and examiners' marking processes.

As reported in Issue 8 of this journal, the findings of Johnson and Nádas (2009) showed that marking GCSE English Literature essays on screen had no significant effect on marker accuracy when compared with how they were marked on paper, although the examiners did exhibit different marking behaviours when marking in each mode.

The examiners in the Johnson and Nádas (2009) project also experienced significantly heightened cognitive workload levels while they marked on screen. The authors concluded that the examiners may have attained similar levels of accuracy across modes because they had sufficient spare cognitive capacity to accommodate the additional cognitive workload exacted by the screen marking task. Based on this conclusion, the authors suggested that the marking accuracy findings may not generalise to extended essays, therefore recommending that further research should explore mode-related effects in the marking of essays with lengths greater than those which were the focus of the earlier studies.

Research questions and research design

To investigate further the potential links between marking mode and the outcomes and processes of extended essay marking, the current project replicated the Johnson and Nádas (2009) project, replacing GCSE essays with longer Advanced GCE essays.

The current project considered six research questions in three broad areas of enquiry, exploring mode-related influences on (i) marking outcomes, (ii) manual marking processes and (iii) cognitive marking processes. The six questions are displayed in Figure 2.

Research Questions

Mode-related influences on marking outcomes were considered through two research questions (RQs):

RQ1: Is examiner marking accuracy influenced by marking mode?

RQ2: Is examiner recognition of essay quality influenced by marking mode?

Mode-related influences on manual marking processes were considered through three research questions:

RQ3: Is examiner manual interaction (i.e. physical contact) with essays influenced by marking mode?

RQ4: Is examiner essay navigation influenced by marking mode?

RQ5: Is examiner annotation practice influenced by marking mode?

Mode-related influences on cognitive marking processes were considered through one research question:

RQ6: Is examiner cognitive workload influenced by marking mode?

Figure 2 : Research questions

This project used an essay question with a maximum of 60 marks available from an Advanced GCE History unit. One hundred and eighty essays from the June 2009 examination session were selected and split into two samples of 90 essays which were broadly similar in terms of mean marks (from the live session) and mark distributions. Table 1 shows the sample features of the essays used in the current project, compared to the sample used in the Johnson and Nádas (2009) project, which used GCSE English Literature essays.

Table 1: GCE History and GCSE English Literature essay sample features

	N	*Written A4 pages*	*Written lines*	*Estimated word count*
		Mean	*Mean*	*Mean*
GCE History project	180	5.3	123.5	900
GCSE English Literature project	180	3.4	75.8	573

The 180 essays were blind marked on paper by the examination's Principal Examiner (PE) to establish a project reference mark for each essay. A sample of 12 Advanced GCE examiners participated in the project. The examiners were all relatively experienced, holding between 6 and 31 total years' experience (mean 16.8 years) of marking for large-scale educational assessment agencies in the UK. Five of the examiners had some previous experience of marking essays on screen.

The 12 examiners marked one of the two samples on paper and the other sample on screen. To control for essay sample and for marking order, a crossover research design was used and the examiners were randomly allocated to one of four examiner marking groups. Table 2 shows the crossover research design used.

Table 2: Examiner marking groups and essay allocation design

Examiner marking group	*1st marking*	*2nd marking*
1	Sample 1 – Paper	Sample 2 – Screen
2	Sample 2 – Paper	Sample 1 – Screen
3	Sample 1 – Screen	Sample 2 – Paper
4	Sample 2 – Screen	Sample 1 – Paper

Prior to marking, all 12 examiners attended a two day meeting to be trained in using the marking software and to standardise their marking in both paper and screen modes. Semi-structured interviews were carried out with each examiner after the marking period had finished, to allow the researchers to probe and check their understanding of the data.

Findings

Mode-related influences on marking outcomes

RQ1: Is examiner marking accuracy influenced by marking mode?

Marking accuracy was defined as the extent of agreement between the examiner marks and the corresponding PE reference marks. Marking accuracy was investigated by considering the differences between the examiners' marks and the reference marks awarded for each essay. These analyses considered two distinct measures of marking accuracy: *absolute*[1] and *actual*[2] mark differences. These measures give an indication of the magnitude and direction of marking accuracy differences between the examiners and the PE for each essay. Descriptive and general linear modelling statistical analyses were then used to investigate whether examiners' marking accuracy was influenced by marking mode.

Table 3 shows descriptive statistics of absolute and actual mark differences between examiner and PE marks by marking mode. Descriptive analyses of absolute mark differences revealed that in both marking modes half of all examiner marks were awarded within five marks of the corresponding PE reference mark. Given the 60-mark range available for the essays, this suggests close equivalence in the overall magnitude of marking accuracy on paper and on screen. Furthermore, a disparity of just 0.08 marks between mean absolute mark differences was identified across modes. Descriptive analyses of actual mark differences add greater depth to this picture. On paper the overall median absolute mark difference was 0 and mean absolute mark difference 0.02, indicating a balance of leniency and severity in marking. In contrast, on screen the overall median absolute mark difference was 1 and mean absolute mark difference 0.47, indicating a very slight tendency towards more lenient marking on screen.

Table 3: Absolute and actual mark differences between examiner and PE marks by marking mode

	Marking mode	
	Paper	*Screen*
N	*1080*	*1067*
Absolute mark difference		
Mean	5.82	5.74
Standard Deviation	4.86	4.45
Median	4.5	5
Actual mark difference		
Mean	0.02	0.47
Standard Deviation	7.59	7.25
Median	0	1

1. The absolute difference between an examiner mark and the corresponding PE reference mark. This measure assigns all differences a positive value, regardless of their direction. Absolute mark differences therefore provide a clear indicator of the *magnitude* of marking accuracy: smaller absolute mark differences represent greater marking accuracy.
2. The actual difference between an examiner mark and the corresponding PE reference mark. This measure assigns a negative value to marks below the reference mark and a positive value to marks above the reference mark. Actual mark differences therefore provide a useful indicator of the *direction* of marking accuracy: negative actual mark differences represent severe marking and positive actual mark differences represent lenient marking.

To enhance the descriptive outcomes, general linear modelling was used to test the statistical significance of any association between marking mode and marking accuracy (Table 4). No statistically significant association between absolute mark differences and marking mode was identified. This reiterated the findings of the descriptive analyses, confirming that there was no statistically significant mode-related difference in the overall magnitude of marking accuracy.

Analyses of actual mark differences suggested a significant association between marking mode and the direction of marking accuracy. Compared to the reference marks, essays marked on screen tended to be marked slightly more leniently than on paper, with screen-marked essays being awarded an average of 0.44 marks more than paper-marked essays. This small difference was statistically significant at the 5% level. Nevertheless, the effect size of this result, another statistical indication of the estimated strength of the relationship, was almost negligible (partial eta squared = 0.002), highlighting an extremely weak association.

Overall, the general linear models found no significant association between marking mode and the magnitude of marking accuracy, and only a small and extremely weak association between marking mode and the direction of marking accuracy. The findings therefore suggest that the examiners were marking with similar accuracy in both marking modes.

Table 4: Results for general linear models of absolute and actual mark differences between examiner and PE marks

		ANCOVA table (N = 2147)					
Variable	*DF*	*Model 1.1: Absolute mark difference*			*Model 1.2: Actual mark difference*		
		Type III SS	*F*	*p*	*Type III SS*	*F*	*p*
Marking mode	1	4.23	0.26	0.61	106.10	4.14	< 0.05
Examiner	11	789.19	4.34	< 0.01	10481.91	37.20	< 0.01
Essay sample	1	61.07	3.70	0.05	3002.49	117.20	< 0.01
Individual essay (nested in essay sample)	1	13453.51	4.57	< 0.01	54497.48	11.95	< 0.01
Error	1955	32308.83			50083.57		

ANCOVA, analysis of covariance; DF, degrees of freedom; SS, sum of squares

RQ2: Is examiner recognition of essay quality influenced by marking mode?

To investigate this question the features which the PE felt were contributing to essay quality were elicited using a modified Kelly's Repertory Grid method (Kelly, 1955). The PE then rated each of the sample essays against each of these essay features to generate a measure of quality for each essay. Finally, these measures were added to the marking accuracy general linear models to investigate whether examiner recognition of essay quality was equal across modes.

The marking accuracy findings from RQ1 indicated that, on average, the examiners marked essays with similar accuracy on screen as on paper. It was not possible to know, however, whether the examiners' recognition of essay quality was also similar across modes (for example, were the examiners better on screen at marking lower quality essays but worse at marking higher quality essays?). When a measure of essay quality was added to the marking accuracy models, analyses showed that examiner recognition of essay quality was not influenced by marking mode. In other words, the examiners marked high and low quality essays with equal accuracy on paper and on screen.

Together, the findings of RQs 1 and 2 support the conclusion that the accuracy of the examiners' extended essay marks and their recognition of essay quality are not influenced by marking mode, and that accurate and valid marking of extended essays is feasible on screen.

Mode-related influences on manual marking processes

RQ3: Is examiner physical interaction with essays influenced by marking mode?

Data about how examiners tangibly interacted with the essays in both modes (e.g. how they physically touched the essays) were gathered through direct observation of one examiner from each of the four marking groups and augmented by interview evidence from all 12 examiners. The observed behaviours were:

- Tagging – physically holding a position in a text while looking at another text to relate two things;
- Overlapping pages in the line of vision;
- Dynamic Tracking – horizontal physical movement with a finger, pencil or mouse during reading;
- Static Tracking – vertical physical movement with a finger, pencil or mouse during reading;
- Pointing/Circling with a focus on one particular aspect (for example, a word) in the text.

The behaviour profiles gathered for the four observed examiners varied in terms of the number and variety of physical interactions that they used on paper and on screen, suggesting that these behaviours reflect highly personalised reading styles.

Overall, the four observed examiners physically interacted less with the essays on screen. Observation evidence suggested that examiners demonstrated fewer focused attention behaviours (i.e. indications that the examiner was attending to a particular word or piece of information; static and dynamic tracking and pointing/circling) on screen, whilst comparative referencing behaviours (i.e. indications that they were attending to more than one piece of information simultaneously; overlapping and tagging) did not alter across modes.

Some evidence from the examiner research interviews suggested that the increased tendency to interact physically with paper was because it was physically and mentally easier to do so in that mode.

RQ4: Is examiner essay navigation influenced by marking mode?

Data for this area of enquiry were also gathered via direct observation of the four examiners and interview evidence from all 12 examiners. The observations captured data about examiners' navigating behaviours while reading essays in both modes, specifically identifying the number of backward reading movements and movements of focus to other documents, such as mark schemes, question papers and other marked essays. Figure 3 shows the mean number of navigating behaviours per observed page by marking mode.

The observation evidence shown in Figure 3 suggests that examiners attended to the mark scheme, question paper or to other marked scripts relatively infrequently whilst marking, with no notable mode-related differences.

In contrast to the observation evidence, however, in the interviews six examiners suggested that they tended not to return to previously marked

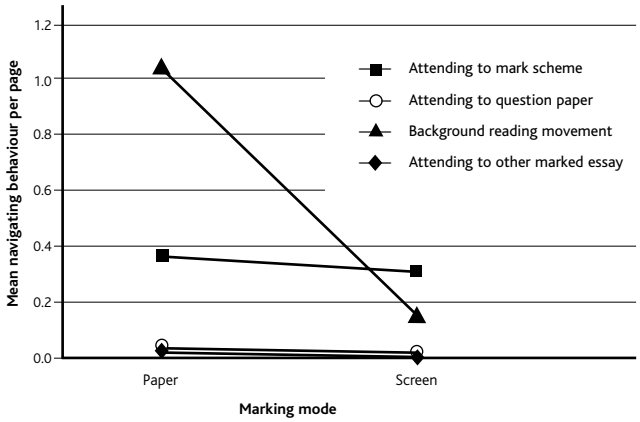


Figure 3 : Mean number of navigating behaviours per observed page by marking mode

essays as readily on screen. Examiners felt that this difference was due to such activity being more difficult to carry out on screen, for example:

> *"Well, I suppose I felt frustrated because it's so difficult…if you wanted to go back three scripts…I thought, 'Oh, can I be bothered with all this clicking and faffing and navigating it, and re-reading it and all this?', and I thought, 'No, I can't'."* (Examiner 8 interview)

Observation evidence also showed that examiners tended to read in a more linear fashion when marking on screen, with fewer iterative or backward reading movements. Examiners suggested in interviews that this was due to the relative difficulty of navigating around essays in this mode:

> *"It's an easier act physically just to turn the page over than to scroll back."* (Examiner 2 interview)

RQ5: Is examiner annotation practice influenced by marking mode?

Thirty essays from essay Sample 1 and 30 matched essays from essay Sample 2 were selected for annotation analysis. The 60 selected essays had each been marked by all 12 examiners and by nature of the research design, each examiner had marked 30 of the selected essays on screen and 30 of the selected essays on paper. Evidence of annotating behaviours was gathered through coded analyses of the marked essays. Again, these data were augmented by interview data from all 12 examiners.

The examiners were able to use a wider variety of annotations on paper than on screen. The screen environment allowed 17 annotation types, including a highlight/underline function. These annotations were built into the marking software following consultation with the examination's PE. For analyses purposes these annotations were termed the 'restricted' annotation palette. Any additional annotations used by examiners when marking on paper were termed the 'extended' annotation palette.

Figure 4 shows the differences in the use of annotations by mode and also by annotation palette. Comparing both the extended and restricted annotation palettes, examiners used an average of 35 annotations per essay on paper and 6 per essay on screen. A Wilcoxon Signed Rank test confirmed that this large mode-related difference was statistically significant ($z = -3.06$, $p < 0.01$, $r = 0.62$). Perhaps this finding is not surprising given that the examiners had access to a limited number of annotation types in the screen marking environment.

When analyses compared only the restricted palette annotations that were available in both marking modes it was found that examiners still annotated less on screen, with a Wilcoxon Signed Rank test confirming this difference to be statistically significant ($z = -2.82$, $p < 0.01$, $r = 0.58$). However, analysis at individual annotation level found that this difference was based on examiners using significantly more underline, sideline and tick annotations on paper. Therefore, when these three annotations were excluded from the overall analysis, there was no significant difference in examiners' use of the remaining restricted palette annotations on paper and on screen.

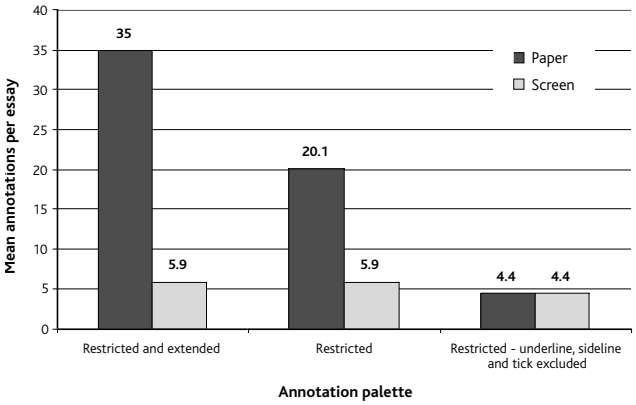


Figure 4 : Mean annotations per essay by marking mode and annotation palette

Examiner interview data were used to help explore the reasons for these mode-related differences. In interviews examiners suggested that they annotated less on screen because the process of using annotations was more difficult and that this might be related to issues of technical usability and their individual levels of proficiency at using the software. Reasons for more limited annotation on screen were also due, in part, to the way that the screen annotation palette sometimes lacked relevance for examiners.

Overall it was evident that physical marking processes were to a large degree idiosyncratic to individual marking behaviours. There was also a clear indication that mode influenced many aspects of examiners' manual marking processes. The physical interaction, navigation, and annotation behaviours that examiners employed for paper-based marking were more difficult for them to replicate when marking on screen.

Mode-related influences on cognitive marking processes

RQ6: Is examiner cognitive workload influenced by marking mode?

Quantitative data about the levels of cognitive workload experienced in each marking mode were gathered using a modified version of the National Aeronautics and Space Administration Task Load Index (NASA TLX) (Hart and Staveland, 1988). The NASA TLX is a self-report survey designed to elicit subjective estimates of the cognitive workload experienced by an individual while performing a specific task. It is underpinned by the assumption that cognitive workload may be represented by a combination of six underlying factors: '*mental demand*', '*physical demand*', '*temporal demand*', '*performance*', '*effort*', and '*frustration*'. The NASA TLX survey was completed twice by 11 of the 12 examiners, midway through their marking sessions in each mode. The survey data enabled a statistical comparison of the cognitive workload

experienced by each examiner across modes to explore whether screen marking was more demanding than paper marking.

Analyses of these data revealed that the examiners experienced greater overall cognitive workload while marking on screen. A Wilcoxon Signed Rank test statistically confirmed that overall cognitive workload was significantly greater on screen ($z = -2.85$, $p < 0.01$, $r = 0.61$). The primary underlying sources of this mode-related difference were identified as the *physical demand* and *fatigue* factors.

Evidence from interview data suggested that the heightened physical demand experienced by the examiners during screen marking was attributed to three key areas of demand: using fine motor skills to operate the computer; maintaining a suitable position at the workstation; and looking at the computer screen. The latter of these physical demands, looking at the computer screen, was highlighted as the most common cause of the fatigue experienced by examiners whilst marking on screen. However, examiner interview comments suggested that this reflected their lack of familiarity with the marking software and might be expected to diminish as their experience of the marking software grows.

Discussion

This project sought to investigate the feasibility of marking extended essays on screen by exploring the potential links between marking mode, essay marking outcomes and marking processes in three broad areas of enquiry;

(i) marking outcomes,

(ii) manual marking processes, and

(iii) cognitive marking processes.

It should be noted that the generalisability of the project findings might be limited by several factors. As a marking simulation exercise, the project differed from a true live marking session in the following key ways:

- The outcomes of the marking exercise had no consequence for candidates, which may have affected examiners' sense of responsibility.
- The marking exercise afforded a comparatively generous time allowance.
- The total marking allocation of 180 essays was comparatively light.
- The previous marking experience of the participating examiners was relatively high.

Marking outcomes

This investigation aimed to consider whether examiners awarded marks which were equally close to the 'true' essay marks in both marking modes. Findings from the statistical analyses suggested that there was no mode-related influence on the magnitude of examiner marking accuracy, but a significant association between marking mode and the direction of examiner marking accuracy was identified. Screen-marked essays were, on average, awarded 0.44 marks more than paper-marked essays. However, the effect size of this result indicated an extremely weak association, and in the context of a 60-mark range the importance of less than half a mark difference is certainly debatable. In light of these perspectives, the findings presented no substantial evidence to indicate that overall marking accuracy was influenced by marking mode.

The examiners' recognition of essay quality across marking modes was also explored. Findings from the statistical analyses suggested that there was no mode-related influence on examiner recognition of essay quality. The examiners attended equally to essay quality when they marked in both marking modes, and the marks awarded recognised that quality.

Together, the marking outcomes findings support the conclusion that the accuracy of the examiners' extended essay marks and their recognition of essay quality are not influenced by marking mode, and that accurate and valid marking of extended essays is feasible on screen.

Manual marking processes

When analyses shifted from marking outcomes to manual marking processes, mode-related influences became more pronounced. The examiners' manual marking processes were broken down into three separate processes: physical interaction, navigation, and annotation. Mode appeared to have an influence on all three of these processes.

The findings show that overall, the examiners physically interacted with essays less on screen than on paper, demonstrating fewer focused attention behaviours when marking on screen. The data did suggest, however, that examiners' physical interaction behaviours were highly personalised, varying widely across individual examiners. Again, when looking at evidence about navigation both within and across essays there were pronounced mode-related tendencies. Evidence showed that the examiners tended to navigate less iteratively on screen and read the essays in a more linear fashion. The most commonly articulated explanation for this difference was the relative difficulty of carrying out traditional paper-based navigation processes on screen.

The examiners in this study also used fewer annotations when marking on screen, due in part to the limited annotation palette available to them on screen. Although the examiners were trained in the use of the software annotation tools it was clear that the examiners still felt that the process of using annotations for marking on screen was too burdensome.

Despite these mode-related differences, examiners were still able to mark extended essays on screen with similar accuracy levels to their paper marking. This implies that the changes in manual marking processes induced by the shift in marking mode did not influence their marking outcomes.

Cognitive marking processes

The examiners experienced greater cognitive workload when marking on screen and this was due to two particular factors – physical demand and fatigue. The examiners attributed the heightened physical demand during on screen marking to the use of fine motor skills to operate the computer, maintaining a suitable position at the workstation or looking at the computer screen. Looking at the computer screen was also highlighted as a common cause of increased and more rapidly arising fatigue.

It is possible that there is an inherent cognitive workload needed when long-held working practices are changed and individuals have to accommodate new ones. The screen marking software influenced examiners' marking processes and these changes could have been initially challenging for the examiners, requiring greater effort. Some of the heightened workload experienced by the examiners could be attributed to their lack of familiarity with the screen marking software, and therefore it is possible that the difference between cognitive workload levels reported across modes might be reduced as examiners' screen marking experience increases.

Conclusion

Returning to the theorised links between extended essay marking mode, processes and outcomes (Figure 1), it appears that mode does have an important influence on some examiner marking processes, but that this does not necessarily influence their marking outcomes. The key practical implication of the findings of this project is that extended essays can be marked on screen without necessarily compromising accuracy. This project supports the conclusions of the Johnson and Nádas (2009) project, and quantitatively demonstrates that the marking of extended essays on screen is feasible. The finding that mode did not present a systematic influence on essay marking outcomes can help to reinforce the defensibility of those marking outcomes and contributes in some way to the maintenance of levels of trust in the assessment system. These findings are of great importance to educational assessment agencies and their stakeholders, and potentially opens the way to the expansion of screen marking to high stakes assessments involving extended essays.

References

Coniam, D. (2009). A comparison of onscreen and paper-based marking in the Hong Kong public examination system. *Educational Research and Evaluation*, **15**, 3, 243–263.

Dillon, A. (1994). *Designing usable electronic text*. London: Taylor & Francis.

Fowles, D. (2008). *Does marking images of essays on screen retain marker confidence and reliability?* Paper presented at the International Association for Educational Assessment Annual Conference, 7–12 September, Cambridge, UK.

Hart, S.G. & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: P.A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*. Amsterdam: North Holland Press, 239–250.

Johnson, M. & Nádas, R. (2009) An investigation into marker reliability and some qualitative aspects of on screen marking. *Research Matters: A Cambridge Assessment Publication*, **8**, 2–7.

Just, M.A. & Carpenter, P.A. (1987). *The psychology of reading and language comprehension*. Boston: Allyn and Bacon.

Kelly, G.A. (1955). *The Psychology of Personal Constructs*. New York: Norton.

Marshall, C.C. & Bly, S. (2005). *Turning the page on navigation*. Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, 7–11 June 2005, Denver, Colorado, USA, 225–234.

Mayes, D.K., Sims, V.K. & Koonce, J.M. (2001). Comprehension and workload differences for VDT and paper-based reading. *International Journal of Industrial Ergonomics*, **28**, 367–378.

O'Hara, K. & Sellen, A. (1997). *A comparison of reading paper and on-line documents*. In: Proceedings of the ACM Conference on Human Factors in Computing Systems, Atlanta, Georgia. ACM Press, New York, 335–342.

Piolat, A., Roussey, J-Y. & Thunin, O. (1997). Effects of screen presentation on text reading and revising. *International Journal of Human-Computer Studies*, **47**, 565–589.

Rose, E. (2010). The phenomenology of on-screen reading: university students' lived experiences of digitised text. *British Journal of Education Technology*. DOI: 10.1111/j.1467–8535.2009.01043.x

Shaw, S. & Imam, H. (2008). *On-screen essay marking reliability: towards an understanding of marker assessment behaviour*. Paper presented at the International Association for Educational Assessment Annual Conference, 7–12 September, Cambridge, UK.

Wästlund, E., Reinikka, H., Norlander, T. & Archer, T. (2005). Effects of VDT and paper presentation on consumption and production of information: psychological and physiological factors. *Computers in Human Behavior*, **21**, 377–394.

EXAMINATIONS RESEARCH

The effects of GCSE modularisation: a comparison between modular and linear examinations in secondary education

Carmen L. Vidal Rodeiro and Rita Nádas Research Division

In this article, a summary of some key aspects and findings from a research project carried out to investigate the effects of modular assessment at GCSE level is presented. The research is described in depth in Vidal Rodeiro and Nádas (2010).

Introduction

GCSEs (General Certificates of Secondary Education) are the qualifications taken by the largest number of students in England. Over five million GCSEs were awarded in 2009, across a range of more than 40 subjects.

As part of the reform of 14–19 education, the national regulator in England revised the subject criteria for GCSEs in collaboration with teachers, awarding bodies, subject associations, higher education organisations and other interested parties. One of the main changes to these qualifications was the increase in the number of unitised or modular specifications.

Up to 2008, modular GCSE specifications were mainly confined to English, ICT, mathematics and science subjects, but since September 2009 almost all specifications are modular in structure, meaning that GCSEs are more in line with A levels, which have been modular since 2000.

A modular specification is one in which the content is divided into a number of units or modules, each of which is examined separately. Module examinations may be taken in different sessions (e.g. January, March, June) and any or all modules may be retaken if the student wishes, with the highest mark for each module retained. However, GCSE

qualification criteria (QCA, 2008) states that unitised specifications must:

- contain a maximum of four assessment units in a single award;
- allocate a weighting of at least 20% to each assessment unit;
- allocate a weighting of at least 40% to terminal assessment;
- allow only one re-sit of an assessment unit, with the better result counting towards the qualification (subject to the terminal rule);
- ensure results for a unit have a shelf-life limited by the shelf-life of the relevant specification.

Linear specifications are usually examined after two years of continuous study, and a candidate normally sits two, three or four papers.

The OCR[1] awarding body took this opportunity to improve the quality of their GCSEs in three key areas: updated and relevant content; focus on developing students' personal, learning and thinking skills; and flexible assessment.

This research focussed on the third key area: *flexible assessment*. This change, which was developed by OCR following extensive consultation (involving teachers, heads of department, local authority advisers, subject associations, professional membership groups and other subject experts), gives schools the flexibility to choose the assessment approach best suited to their students.

The assessment of the new OCR GCSEs is organised into units which can either all be taken at the end of the course in a linear fashion, or can be taken in different sessions throughout the course (for many subjects, assessment is available twice a year) to follow a more unitised approach to teaching and learning. It should be borne in mind that unitised does not mean staged. Units can be taken in any order, rather than being restricted to being assessed in a particular sequence.

When modular and linear paths exist for the same subject, it is left to the schools to decide whether the assessment of any particular subject should be modular or whether they should enter candidates for the linear examination.

Over recent years, there has been a clear trend in the development of the upper secondary curriculum to increase the use of modular or unitised qualifications. In particular, in the 1980s much interest was shown in modular courses and many such courses were developed and introduced in British secondary schools. As a result, the rationale for modularisation and many of the issues arising from it were addressed (see, for example, SEC, 1987; Moon, 1988; or Warwick, 1987).

The drive behind some of the attempts to modularise qualifications came from teachers seeking to make the curriculum more relevant to their students and to increase their motivation through the setting of short-term assessment targets. An early example of modular assessment within the school examination system is described in Thomas (1993), who discussed the introduction of a modular science course and the reactions of teachers to this course, focussing, in particular, on the impact on organisational issues and teaching methodology.

The earliest attempts to modularise A levels occurred in the 1980s/1990s, for example, the Wessex A levels (Macfarlane, 1992) or the UCLES scheme (UCLES, 1986; Nickson, 1994). However, by the early 1990s there were already concerns about modular courses being too easy in comparison with terminally examined courses. Some of the reasons for these concerns were: modular courses had been associated with lower attaining students; candidates could retake modules to improve grades; and candidates could be examined on parts of a subject rather than on the entire syllabus. Others argued that modularisation could make the courses more difficult because candidates were expected to work and be assessed at A level standard from the first module taken early in the first year and therefore might be potentially disadvantaged by their relative immaturity, if not their narrower experience of the subject. In fact, on the subject of modular syllabuses, UCAS (1994) stated 'It should be clearly understood that modular syllabuses are no easy option, as all modules are assessed to full GCE A level standard without allowance for maturation, including those taken at an early stage in the course'. It was the Dearing Review (Dearing, 1996) that provided the template for the current model of modular A levels and led to the development of the 'Curriculum 2000'. As a result of the implementation of this initiative, a number of evaluations and reviews were carried out to ensure the validity and reliability of the modular assessment and the challenges to the quality of teaching and learning (e.g. Hayward and McNicholl, 2007).

The proponents of modular schemes have long argued for their advantages in terms of curriculum flexibility, short-term assessment goals, regular feedback, re-sit opportunities and increasing motivation for students. On the other hand, critics of the modular assessment claim that it leads to fragmentation of learning, students entering examinations when not ready, more teaching to the test and over-assessment. Furthermore, it is also being claimed that GCSEs are becoming less and less demanding, which might lead to a diminution of trust in the qualification among the general public, higher education tutors and university admissions staff. Studies have identified a number of advantages and disadvantages of modular assessment which are discussed in detail in Vidal Rodeiro and Nádas (2010) and briefly outlined below.

Advantages of modular courses and assessment

- There is a choice of learning approach: linear or unitised;
- the assessment can be timed to match the point of learning within the course, making it easier for candidates to show what they know, understand and can do;
- students can re-sit a unit rather than repeat the entire assessment;
- modular feedback enables students to 'remedy weaknesses' before the final examination;
- students are better motivated as they receive feedback on performance more frequently and earlier in the course;
- the pace of students' work is brisk at the beginning of the course;
- a unitised approach makes it easier for students to stay on track with their studies and manage their time effectively;
- the assessment load is spread more evenly over two years, reducing examination stress and the pressure of an 'all or nothing' assessment;
- modular assessment enables students to plan their studies;
- revision is more manageable;
- assessment is potentially more reliable because it is based on more assessed work in total (e.g. in GCSE in Religious Studies there were 147 raw marks available in the linear specification (OCR, 2000) and 192 in the modular specification (OCR, 2009));
- with a similar format to A levels and Diplomas, the unitised GCSEs will help prepare students for the next phase of their education;
- a sense of ownership is forged, leading to less disaffection among students.

1. Oxford Cambridge and RSA Examinations

Disadvantages of modular courses and assessment

- There is a danger of fragmentation of learning and lack of coherence in learning programmes due to both the teaching methods and the assessment practices;
- there can be a poorly developed overview of subjects and an inability to connect discrete areas of knowledge;
- adopting a modular approach can disrupt the provision of a coherent and developmental course;
- assessment becomes dominant throughout the course, rather than towards the end of it;
- deadlines on units can limit a teacher's ability to teach important topics in the way that he or she would choose;
- it is possible for a student to sit an examination before being ready (disregard for individual intellectual maturity);
- short-term targets often dominate over longer-term goals, encouraging a cram-and-discard approach;
- if re-sits are not well managed, students could re-sit too many modules. This increases pressure on school resources and on students' workload;
- the general public, higher education tutors and admissions staff are less trusting of modular qualifications, which they perceive to be easier.

Aim of the research

This project combined quantitative and qualitative research methods to investigate the impact of modular assessment on GCSE students.

The main aim of the statistical strand was to explore the differences in outcomes between candidates who took assessments in GCSE specifications in a terminal or linear approach (all units at the end) and those who adopted a modular approach (taking units throughout the two-year course).

The qualitative strand of the project aimed to investigate, in the school context, the effects of modularisation on students and teachers in terms of motivation, consistency and amount of workload, exam pressure and effects of feedback.

In particular, the research aimed to answer the following questions:

1. Are there differences in examination outcomes between the students who take assessments in a terminal or linear approach and those who adopt a modular approach, once their general ability is taken into account?
2. Are students at a disadvantage by their relative immaturity or narrow experience of the subject if they enter for an examination early?
3. Are students benefiting from being able to re-sit modules?
4. Does regular feedback (positive or negative) motivate students?
5. Does regular feedback help students to identify their learning needs?
6. Does modular assessment remove the pressure of an all-or-nothing exam?
7. What are the characteristics of modular students' test-taking motivation?
8. What are teachers' attitudes towards modularisation and what is the impact of modular assessment on their workload?

Research methods

Previous research (e.g. Ofsted, 1999) has suggested that modular specifications work most successfully in subjects such as mathematics or physics and are less suited to subjects like English or modern foreign languages. Therefore, two contrasting subjects at GCSE level were selected for this research: English and mathematics. Only candidates who sat an examination in these subjects with the OCR awarding body were considered.

For the quantitative strand of this research, examination outcomes in both subjects, at specification and at unit level, were obtained from OCR's examinations processing system. The data comprised personal details (name, sex, date of birth and school) and assessment details (session, tier, final mark and final grade). Six successive cohorts of English students (2004–2009) were investigated. However, as the unitised GCSE mathematics specification was first certificated in 2008, only two cohorts of mathematics students (2008–2009) were available for analysis. Descriptive statistics were used to investigate the entries and the re-sit patterns for both assessment routes and regression analyses were carried out to explain the differences in attainment between linear and modular routes once the general ability of the students, measured by prior/concurrent attainment at school, was taken into account.

In the qualitative strand of the research, questionnaires and face-to-face interviews with students and teachers in schools offering either modular or linear GCSE English and/or GCSE mathematics were carried out in order to collect data on motivation, feedback, exam pressure and workload. In particular, data on motivation was collected using an intrinsic motivation inventory survey developed by Ryan and Deci (undated), which has six subscales – choice, competence, effort, enjoyment, pressure and value – that measure different aspects of test-taking motivation; effects of feedback on students were mapped in interviews conducted after candidates had received the grade reports on the unit examinations; and perceived workload data were collected via a survey in the form of a self-report workload chart for students and teachers to fill in retrospectively. In the qualitative strand of the research, 62 students and two teachers of GCSE English (all in one school) and 61 students and two teachers of GCSE mathematics (grouped in two schools) took part.

The structure of the two subjects considered in this research is described briefly below.

GCSE in English

The OCR GCSE in English (OCR, 2003) has a unit-based structure, enabling both linear and modular assessment routes. Units which are externally assessed by written examination contain two options: a foundation tier component and a higher tier component. Coursework units are not tiered. Table 1 shows the specification structure.

In order to certificate for a GCSE in this subject, at least four units must be taken, including:

– one component from Unit 1

– one component from Unit 2

– *either* one component from Unit 3 or Unit 4

– Unit 5

Although the specification is unit-based, it is possible to follow a linear route and take all the necessary units in the same examination session.

For the modular/unitised route, four or more units, as specified above, may be entered across two or more examination sessions. Units may be re-taken once, if wished, prior to certification and the better score will be used towards the overall grade (subject to the terminal rule). However, at least 50% of the qualification needs to be taken as terminal external assessment.

The first certification session for this qualification was June 2004. Thereafter, assessment was available in January and June each year.

Table 1: OCR GCSE in English structure (OCR, 2003)

Unit	*Option*	*Title*	*Format*
1	2431 F	Non-fiction, media and information (Foundation Tier)	Written Exam
	2431 H	Non-fiction, media and information (Higher Tier)	Written Exam
2	2432 F	Different cultures, analysis and argument (Foundation Tier)	Written Exam
	2432 H	Different cultures, analysis and argument (Higher Tier)	Written Exam
3	2433 F	Literary heritage and imaginative writing (Foundation Tier)	Written Exam
	2433 H	Literary heritage and imaginative writing (Higher Tier)	Written Exam
4	2434	Literary heritage and imaginative writing	Coursework
5	2435	Speaking and listening	Coursework

GCSE in mathematics

OCR offers three different routes to obtain a GCSE in mathematics:

– GCSE mathematics A: Linear Assessment

– GCSE mathematics B: Mathematics in Education and Industry

– GCSE mathematics C: Graduated Assessment (unitised)

The focus of this research was on GCSE mathematics A and C. Both subjects are identical in content but different in structure.

The scheme of assessment for the GCSE mathematics A (OCR, 2006a) consists of two tiers, foundation and higher. In each tier, candidates have to sit two examination papers and submit coursework. Candidates wishing to re-sit this qualification must re-sit both written papers at the appropriate level but may carry forward their coursework mark.

The GCSE mathematics C (OCR, 2006b; OCR, 2007) has been divided into a series of ten stages which are graduated in content and level of difficulty. Corresponding to each stage a module test was set. Table 2 shows the qualification structure.

Candidates normally take the course over two years and must enter at least two different module tests. Most modules are available in January, March and June sessions and in most cases they target a pair of grades. All candidates have to take one terminal examination. The tier of entry for the terminal examination determines the overall grades available to the candidate. Candidates may re-sit any module test once prior to certification and the better score is used in the aggregation. After certification, candidates who wish to re-sit must sit at least the terminal paper again, but might carry forward their coursework mark (if applicable) and/or their module test marks.

Both qualifications were first certificated in June 2008. Thereafter, certification was available in January and June each year.

Table 2: OCR GCSE in mathematics structure (OCR, 2007)

Units	*Target grade*
M1	G
M2	G,F
M3	F
M4	F,E
M5	E
M6	D
M7	C
M8	B
M9	A
M10	A*
TF – Terminal Paper (Foundation Tier)	G–F, E–C
TH – Terminal Paper (Higher Tier)	D–C, B–A*

Key findings

Entries and assessment route

- Higher percentages of candidates entering for a GCSE in English followed a linear assessment route than a modular assessment route (e.g. 80% vs. 20% in 2009). However, entries for the modular assessment route were on the increase in the period of study and entries for the linear route were decreasing. On the contrary, the majority of the candidates studying for a GCSE in mathematics followed a modular assessment route (e.g. 63% vs. 38% in 2009).
- In four of the five GCSE English units the majority of candidates took the examination in the terminal session. However, the percentages of candidates sitting units in early sessions had been increasing over time. It can be the case that the more able students are being stretched by completing some modules at an early stage and then progressing to other work. The entries for the remaining unit were well spread throughout the two-year course. In GCSE mathematics, for the majority of the units, less than 20% of the entries were for the terminal session. This shows that, in mathematics, candidates made use of the flexible assessment by getting units out of the way rather than taking them in a narrow window at the end of the two-year course. In particular, the majority of the mathematics students interviewed in this research reported that they welcomed external examinations during the school year.
- Previous research into modular examinations (e.g. Ofsted, 1999), showed that modular syllabuses are more successful in mathematics and are less suited to English, where the assessment can interrupt the teaching of themes that run across more than one module. In this research, the proportions of candidates who took all their module examinations in one session suggest that modular assessment is thought less appropriate for English than for mathematics. The results from the qualitative strand confirm that the students of mathematics were generally in favour of modular assessment and the students of English appreciated some characteristics of the modular assessment but they did not express a strong preference towards modularisation.
- Both strands of this research show that the introduction of the unitised specification in GCSE English did not lead to many changes in the way the subject was taught, studied and assessed, as it mostly continued to be addressed as if it were linear in design. Factors such

as maturity or parallel teaching across modules in English might have led many students to sit the majority of their modules terminally.

- The degree of flexibility in the number and timing of the modular examinations was illustrated in this research by the large number of unit combinations that led to a GCSE in each subject. This proves that modular syllabuses are seen as a method of giving students a degree of choice in syllabus content and assessment session. However, the most frequent combinations of modules may be more likely to reflect the teaching resources available within a centre or the schools' preferences as opposed to any other factors. The reasons why schools offer modular syllabuses in certain subjects or prefer the linear approach in others warrants further study as modular courses are becoming increasingly popular.
- It should be noted that due to the 'newness' of the modular schemes at GCSE, the pattern of entries may be reflecting some experimentation on the part of the teachers in deciding the points in the course when their students should sit the examinations. Also, it is possible that different patterns of entry may emerge as the modular schemes mature and teachers and candidates become more confident in making decisions regarding the most appropriate time to sit module examinations.

Linear assessment vs. modular assessment outcomes

- The quality of the entry in each of the assessment routes was different. GCSE English students following a linear assessment route had, on average, higher ability than candidates following a modular route. Mathematics students following a linear assessment route had slightly lower ability; this might be due to the fact that lower ability mathematics students do not welcome many external exams during the school year due to the additional workload involved and they prefer an end-of-year examination. This fact has been confirmed by the mathematics students interviewed in the qualitative strand of this research. It was important then to take into account students' ability when talking about the performance in each of the assessment routes.
- In GCSE mathematics, candidates following a modular route obtained higher grades than candidates following a linear one once their ability was accounted for. In contrast, and contrary to anecdotal evidence, which suggests that with modular syllabuses it is easier to attain higher grades, modular routes in GCSE English led to lower grades than linear routes.
- It should be noted that the fact that candidates obtain higher grades from a modular scheme does not necessarily mean that standards have dropped. It has been suggested (e.g. Gray, 2001) that, in a modular scheme, setting targets throughout the course, having ongoing feedback and allowing a certain amount of re-taking within the course leads to candidates learning more – thereby obtaining higher grades.

Maturational effects

- According to previous research, candidates cannot be expected to perform as well in early sittings as they would later on in the course (Clarke, 1996; Taverner and Wright, 1997). Students might be at a disadvantage if they are entered for an examination before being ready as they might not have the experience of the two-year course and might be at different levels of age and maturity. Therefore, there can be powerful arguments for linear assessments as certain skills may develop progressively through several modules.
- This research showed that, in the modular routes of GCSE English, students opting for certificating midway throughout the course[2] were at a disadvantage compared to those who opted for certificating at the end. Girls were at a greater disadvantage than boys. The gender effect was in line with previous research which showed that boys were more likely to take advantage of modular examinations than girls (McClune, 2001). On the other hand, girls following a linear assessment route and certificating early in the two year course had a higher probability of achieving a given grade or above than those who certificated late. In English, subject maturity, which is thought to improve performance, is important and the modular route is, therefore, a more difficult one. This finding is supported by previous research (e.g. SCAA, 1996).
- For GCSE English, maturational effects differed by unit. In the modular assessment route, candidates sitting early any of the three externally assessed units (by written examination) did not perform as well as those sitting them later. At unit level, analyses by gender did not reveal statistically significant differences between boys and girls. In the linear assessment route, girls, who are generally considered to mature earlier than boys, seemed more likely than boys to benefit from taking the examination early in any of the three externally assessed units. Boys, on the other hand, seem more likely to benefit from taking the examination in the later part of the course. However, early assessment seemed to be an advantage for both girls and boys in the coursework units in both the linear and the modular routes. Students might have wanted to carry out their coursework assignments early in the course to relieve the workload towards the end of the year and they worked hard to do so.
- GCSE mathematics students obtained, on average, significantly higher marks in early sessions than in later sessions. In particular, candidates taking modules targeting grades A*–D (modules M6 to M10) performed much better, after allowing for their ability, in the earlier sessions than in the terminal session.

Patterns and impact of re-sits

- In both GCSE English and GCSE mathematics the research showed an increase over the period of study in the percentage of students taking re-sits (e.g. the percentages of students taking re-sits in English increased from 4% in 2004 to 10% in 2009 and in mathematics from 46% in 2008 to 52% in 2009). However, some schools have the view that the number of re-sits should be limited since they are expensive, cause timetabling problems and many students do not make sufficient progress to warrant them.
- Looking at the changes in marks/grades between the first and second attempts of a unit, the benefits of re-sitting seem clear. Across all units investigated in this research, the majority of candidates did better on their second attempt than they had on their first, with percentages of students obtaining an improvement in the unit grades ranging from 25% to 65%, depending on the unit and the subject[3].

2. Students completed the course following a modular route but they did so in less than two years.
3. These are percentages of students taking re-sits and not percentages of the total entry in the relevant unit.

It should be borne in mind, however, that the knowledge that a re-sit was available may have lessened a candidates' resolve to do their best at the first attempt. Students of modular syllabuses interviewed in the qualitative strand of this research mentioned that the possibility to re-sit a module relieved some of the stress and pressure of the modular exams and admitted that they would have worked more had there been only one chance for them to pass their examinations.

- The fact that a relatively high percentage of students improved their unit marks/grades after a re-sit taken later in the course may suggest that some students were entered for unit examinations before they were ready. Teachers, therefore, will need to make sure that their students are ready when deciding the points in the course when they should sit the examination. There might be the case that candidates take examinations at an early stage of the course to familiarise themselves with the demands of the modular examinations or as confidence/motivation building sessions. Other candidates might take them at a later stage to improve an earlier result.
- The differences in the re-sitting patterns by centre type were small, with the percentage of students taking no re-sits being higher in the independent sector. This was in line with a study carried out by QCA (2007) about re-sitting patterns and policies in respect to GCE A levels in seven subjects (including English literature and mathematics) which indicated that there was very little difference in the scale of re-sitting behaviour in terms of centre type. However, the QCA study highlighted that there were differences across the different centre types in terms of the training that a candidate might receive when preparing for a re-sit. For example, in a number of independent centres unlimited support had been given to candidates wishing to re-sit in comparison to the majority of the state schools, where past papers tended to be all that was offered to re-sitting candidates.
- Opinions are divided as to whether re-sits should be allowed. Re-sits are perceived by some as unfair as some candidates might not have the opportunity to attempt one unit twice (maybe due to school policies on re-sits or cost). It should be borne in mind that there is some improvement that is 'valid'. For example, students might have performed better in the re-sit than in the first attempt of an examination due to extra teaching or personal circumstances out of their control which may have affected performance at the first sitting. Also, there is a maturation benefit and, for example, students may be able to improve their general understanding and ability in a subject over time.

Regular feedback

- Students of all abilities taking GCSE English or GCSE mathematics found feedback (positive and negative) useful and motivating and reported that it encouraged them to do better on the next modules and/or on the terminal papers. Students appreciated seeing the units' grades and they felt that they received feedback soon enough after sitting the exam. Students also found it useful to be informed about how much improvement they could expect in their terminal paper.
- Mathematics students were more satisfied with their grade reports (most common form of feedback) and gained more information from them than students of English. Furthermore, mathematics students found it easier than English students to identify the strengths and weaknesses of their performances.
- However, grade reports were not helpful in identifying students' learning needs and informing their learning strategies. Students reported missing the opportunity of going through their own marked papers (as the scripts arrived too late after the examination) or receiving suggestions about the areas they needed to improve on in order to change, if necessary, their focus of learning and strategies of exam preparation.

The pressure of an all-or-nothing exam

- Modular assessment does not remove the stress and workload of an all-or-nothing exam.
- Students in the modular routes reported that the pressure to achieve a good grade placed significant stress on them during both the modular and the end-of-year examinations. However, the possibility to re-sit modular examinations was mentioned as helpful in alleviating some of the examination stress experienced during modular exams, as it gave some students confidence about what to expect on their subsequent exams.
- Students of modular mathematics experienced longer periods of higher workload than linear students did in the first half of the year.
- For students of English, the workload varied considerably during the course of the year but there were no differences in linear and modular students' workload levels. Some students found the January modular exams quite stressful due to them coinciding with other unit examinations and the coursework assignments.

Students' motivation on modular exams

- Modular mathematics students perceived their modular exams to be quite valuable, and they were generally motivated to do well. However, the results of the survey indicate that these students did not really 'own' the examination, and that instead of being intrinsically motivated, they perceived it as an externally imposed, compulsory task. Students scored high on pressure and they reported putting forth a lot of effort during the sitting of the examination. On the contrary, students scored very low on perceived competence and perceived choice in sitting the exam, and they obtained the lowest ratings in the enjoyment scale.
- Students of English had high scores on effort and value, which implies that they appreciate the usefulness of the examinations and make appropriate effort to do well on them. Despite feeling under less pressure than mathematics students, students of English had low ratings for intrinsic motivation (i.e. for enjoyment, competence and choice).

Teachers' workload and attitudes towards modularisation

- Teachers in the modular assessment system appreciated the better planning opportunity around the exams, the clarity of the focus of their teaching requirements and felt that modular assessment contributed to their approach to assessment for learning. They also appreciated not having to re-motivate students at the end of the year and felt that modular assessment helped to encourage continuous study and revision in students who were difficult to motivate.
- Teachers in the linear route appreciated having more space and control to deliver the content effectively; furthermore, they did not

find it a burden to revisit topics and re-motivate students before the end-of-year examination. In particular, one teacher of mathematics was concerned about modular students having to revisit materials from long-forgotten modules before the final examinations and felt that the linear route allowed her to deliver the content more effectively and in a more enjoyable and mixed structure.

- Mathematics teachers' workload levels varied with the assessment route: the linear assessment placed very high levels of workload on the teachers at certain times whilst the modular assessment provided a more evenly spread workload rising throughout the year.
- English teachers' workload levels were continually increasing between September and December, when teachers were marking mock exams and preparing for unit examinations in January. From that point onwards, workload levels varied by teacher.

This research has addressed some of the key issues relating to the effects of unitised specifications at GCSE level (e.g. curriculum flexibility, short-term assessment goals, maturity, regular feedback to students, re-sits, increasing motivation) and provides evidence of students' and teachers' general attitudes to modularisation and reasons for the differences in the outcomes of students who took different assessments routes (linear vs. modular). It should be noted though, that the qualitative strand investigated only the views of a selected few students and teachers at three schools who do not represent all the population. However, by reporting the students' voice, the results of the statistical strand were enriched.

References

Clarke, J. (1996). Modularising A levels in social sciences: some conclusions. *Social Science Teacher*, **26**, 1, 33–35.

Dearing, R. (1996). *Review of qualifications for 16–19 years old: full report*. Middlesex: SCAA Publications.

Gray, E.A. (2001). *The comparability between modular and non-modular examinations at GCE Advanced Level*. PhD Thesis. London: Institute of Education.

Hayward, G. & McNicholl, J. (2007). Modular mayhem? A case study of the development of the A level science curriculum in England. *Assessment in Education*, **14**, 3, 335–351.

Macfarlane, E. (1992). *Education 16–19: in transition*. London: Routledge.

McClune, B. (2001). Modular A levels – who are the winners and losers? A comparison of lower-sixth and upper-sixth students' performance in linear and modular A level physics examinations. *Educational Research*, **43**, 1, 79–89.

Moon, B. (1988). *Modular curriculum*. London: Paul Chapman Publishing.

Nickson, M. (1994). *Evaluation of module bank system*. Cambridge: University of Cambridge Local Examinations Syndicate.

OCR (2003). *OCR GCSE in English (Opening Minds) 1900*. Cambridge: Oxford, Cambridge and RSA Examinations.

OCR (2006a). *OCR GCSE in Mathematics A (Linear Assessment) J512*. Cambridge: Oxford, Cambridge and RSA Examinations.

OCR (2006b). *OCR GCSE in Mathematics C (Graduated Assessment) J516*. Cambridge: Oxford, Cambridge and RSA Examinations.

OCR (2007). *OCR GCSE in Mathematics C (Graduated Assessment) J517*. Cambridge: Oxford, Cambridge and RSA Examinations.

OCR (2000). *OCR GCSE in Religious Studies B (Philosophy and Ethics) 1931*. Cambridge: Oxford, Cambridge and RSA Examinations.

OCR (2009). *OCR GCSE in Religious Studies B (Philosophy and/or Applied Ethics) J621*. Cambridge: Oxford, Cambridge and RSA Examinations.

Ofsted (1999). *Modular GCSE AS and A level examinations 1996–98*. London: Office for Standards in Education.

QCA (2007). *A level re-sitting: summary of research findings*. London: Qualifications and Curriculum Authority.

QCA (2008). *GCSE Qualification criteria. For first teaching from September 2009*. London: Qualifications and Curriculum Authority.

Ryan, R.M. & Deci, E.L. (undated). Intrinsic Motivation Inventory. Available online at: http://www.psych.rochester.edu/SDT.

SCAA (1996). *Evaluation of modular A levels*. London: School Curriculum and Assessment Authority.

SEC (1987). *Assessing modular syllabuses in the GCSE and at GCE advance and advance supplementary levels. A discussion document*. London: Secondary Examinations Council.

Taverner, S. & Wright, M. (1997). Why go modular. A review of modular A level Mathematics. *Educational Research*, **39**, 1, 104–112.

Thomas, G. (1993). Some reactions to teaching of science using a modular scheme. *Educational Review*, **45**, 3, 213–225.

UCAS (1994). *Examinations and grades*. Cheltenham: Universities and Colleges Admissions Service.

UCLES (1986). *GCE Advanced level module bank*. Cambridge: University of Cambridge Local Examinations Syndicate.

Vidal Rodeiro, C.L. & Nádas, R. (2010). *Effects of modularisation*. Research Report. Cambridge: Cambridge Assessment.

Warwick, D. (1987). *The modular curriculum*. Oxford: Basil Blackwell Ltd.

Tracing the evolution of validity in educational measurement: past issues and contemporary challenges

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

Introduction

Validity is not a simple concept in the context of educational measurement. Measuring the traits or attributes that a student has learnt during a course is not like measuring an objective property such as length or weight; measuring educational achievement is less direct. Yet, educational outcomes can have high stakes in terms of consequences (e.g. affecting access to further education), thus the validity of assessments is highly important.

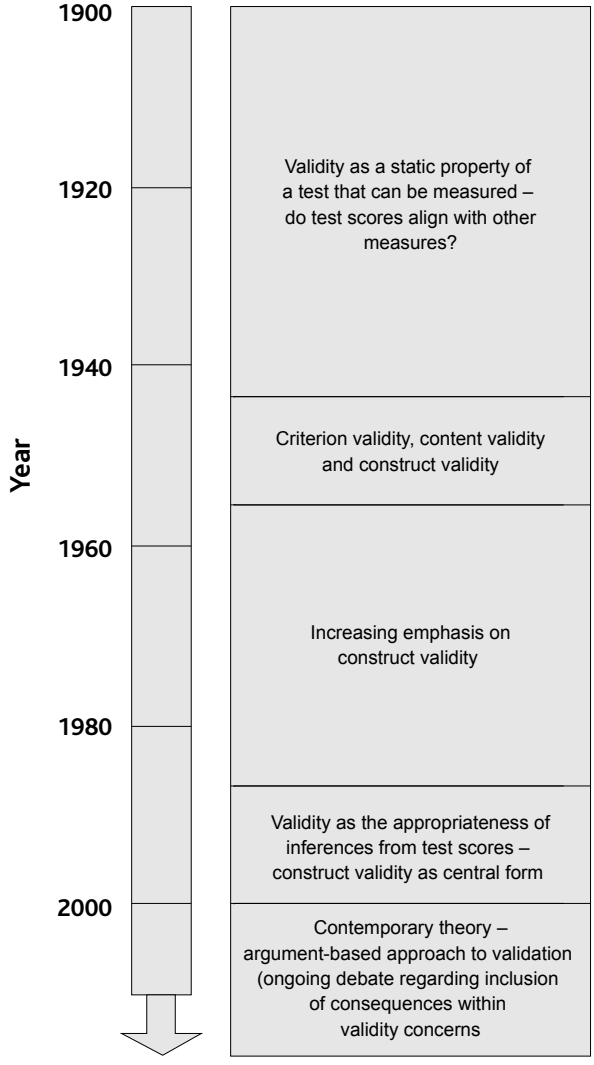


Figure 1 : Timeline of the evolution of validity theory

Tracing this trajectory of evolution, particularly through key documents such as the validity/validation chapter in editions of Educational Measurement (Cureton, 1951; Cronbach, 1971; Messick, 1989; Kane, 2006) and the Standards of Educational and Psychological Testing (AERA, APA and NCME, 1954/1955, 1966, 1974, 1985, 1999) has been important to us as part of work to develop an approach to validation for general assessments.

The concept of validity is not a new one. Conceptualisations of validity are apparent in the literature from around the turn of the twentieth century, and since that time, they have evolved significantly. Earliest perceptions of validity were that of a static property captured by a single statistic, usually an index of the correlation of test scores with some criterion (Binet, 1905; Pearson, 1896; Binet and Henri, 1899; Spearman, 1904). Through various re-conceptualisations, contemporary validity theory generally sees validity as about the appropriateness of the inferences and uses made from assessment outcomes, including some consideration of the consequences of test score use. This article traces the progress and changes in the theorisation of validity over time and the issues that led to these changes. A timeline of the evolution of validity is provided by Figure 1.

1900–1950: Early validity theory

Most early validity theory was located within a realist philosophy of science[1] and in terms of educational measurement couched within the highly scientific discourse of psychological testing, grounded as it was in a positivistic epistemology. During this time validity was conceived of as a statistical index, validity being evaluated in terms of how well the test scores predicted (or estimated) the criterion scores. The criterion measure was the value (or amount) of the attribute of interest. The attribute was assumed to have a definite value for each person and the objective of assessment was to arrive at an accurate estimation of the amount of attribute manifested. Thus validity was defined in terms of the accuracy of the estimate and validation was seen to require some criterion measure which was assumed to provide the 'real' value of the attribute of interest.

Early definitions placed emphasis on the test itself. Bingham defined validity from an operational perspective as the correlation of scores on a test with "some other objective measure of that which the test is used to measure" (Bingham, 1937, p.214) – a view shared by a number of well known measurement theorists at the time (including Cureton, 1951; Gulliksen, 1950) and most notably expressed by Guilford (1946), who said that "in a very general sense, a test is valid for anything with which it correlates" (p.429).

By the 1920s, tests were described as being valid for any criterion for which they provided accurate estimates (Thorndike, 1918; Bingham, 1937). For example, Kelley noted "the problem of validity is that of whether a test really measures what it purports to measure" (1927, p.14).This view prevailed throughout the first half of the twentieth century.

1. Scientific realism was developed largely as a reaction to logical positivism. Scientific realists claim that science aims at truth and that scientific theories should be regarded as true (or at least approximately true, or likely to be true).

1950s: Criterion-based, content-based and construct-based models of validity

During the 1950s, the concept of validity was refined to include the ability of a test to predict future performance with respect to external criteria (*criterion*), content area (*content*), or a theoretical construct (*construct*). In other words, validity was conceptualised as triune in nature comprising criterion, content and construct facets. Throughout this time, and even beyond, the tripartite division had become so widely embraced that Guion, writing in the 1980s, criticised how many took this structure 'on faith' and without questioning. He referred to it as "something of a holy trinity representing three different roads to psychometric salvation" (1980, p.386).

Criterion validity

The 1950s began with Cureton's sophisticated summary of conceptions of validity which he articulated prior to the advent of construct validity. Cureton (1951) stated that "The essential question of test validity is how well a test does the job it is employed to do" (p.621). Validity, he argued, "indicates how well the test serves the purpose for which it is use[d]" and, therefore, can be "defined in terms of correlation between the actual test scores and the 'true' criterion scores" (p.623). Essentially, he was arguing for the criterion model as offering the best solution to evaluating validity. This view was predicated on earlier conceptualisations of validity as a static property that could be measured in relation to a true criterion.

The criterion-based model, in which validity of the criterion and test scores were to be validated against the criterion scores, was helpful in a variety of applied scenarios, assuming that some suitable 'criterion' measure was available. Apart from being an objective measure, criterion-related evidence seemed relevant to the plausibility of proposed test score interpretations and uses.

In the 1954/1955 *Standards* (AERA, APA and NCME, 1954/1955) criterion validity was deconstructed into two forms of validity: concurrent validity and predictive validity. Concurrent validity made use of indirect measures which permitted validity estimates to be obtained concurrently with test scores, whilst predictive validity depended on a criterion of subsequent performance which could not be achieved concurrently with test scores. The 1966 *Standards* (AERA, APA and NCME, 1966) characterised criterion validity in the following way: criterion validity compared test scores with "one or more external variables considered to provide a direct measure of the characteristic or behaviour in question" (p.12).

However, there were issues with the criterion-based model which demanded a well-articulated and demonstrably valid criterion measure. Presupposing a criterion measure was available, questions about the validity of the criterion emerged. Unfortunately, the model was unable to provide a sound footing for validating the criterion. One possible solution was to employ a criterion measure involving some desired performance and then to interpret the scores in relation to that performance such that validity of the criterion could be accepted.

Content validity

Content validity methods focus on item content and the degree to which the test samples the 'universe' of relevant content. According to the 1966 *Standards* (AERA, APA and NCME, 1966), content validity demonstrated how well a test "samples the class of situations or subject matter about which conclusions are to be drawn." Much later, Messick (1989) described content-validity evidence as providing support for "the domain relevance and representativeness of the test instrument" (1989, p.17). It was deemed legitimate to extrapolate from an observed performance on a sample of assessment tasks from a domain as an estimation of generalised performance in the domain providing it could be demonstrated that the observed performances were representative of all assessment tasks and that the size of the sample was sufficiently large to control for sampling error (Guion, 1977).

However, content-validity evidence tended to be both subjective and confirmatory (based on judgement by experts who sometimes had a vested interest in the assessment) and did not involve test scores or performances on which scores were based. Consequently, it was difficult to justify conclusions about interpretation of test scores. Additionally, the content-based validity model proved to be problematic when used as grounds for arguing the validity of claims about cognitive processes or underlying theoretical constructs as the following quotes illustrate:

- "Judgments about content validity should be restricted to the operational, externally observable side of testing. Judgments about the subjects' internal processes state hypotheses, and these require empirical construct validation." (Cronbach, 1971, p.452)
- Content-based validity evidence provides "the domain relevance and representativeness of the test instrument" (Messick, 1989, p.17) but does not provide direct evidence for the "inferences to be made from test scores" (p.17).

It was becoming increasingly more necessary, given the shortcomings of both the criterion-based and content-based models, to develop a more sophisticated conceptualisation of validity.

Construct validity

Meehl and Challman (APA, 1954) first introduced the concept and terminology of construct validity, however, the concept was developed further by Cronbach and Meehl's (1955) seminal paper – 'Construct validity in psychological tests' – published in *Psychological Bulletin*. Much of their thinking had its origins in the hypothetico-deductive (HD) model of scientific theories (Suppe, 1977). Cronbach and Meehl began with the notion of a construct as "some postulated attribute of people assumed to be reflected in test performance" (1955, p.283), and asked the question whether the test was an adequate measure of the construct. According to Cronbach and Meehl, "determining what psychological constructs account for test performance is desirable for almost any test" (1955, p.282). They suggested that construct validity was an all-pervasive concern though they did not offer it as a general organising framework for validity. Cronbach and Meehl (1955) attempted to link theory and observation – a central tenet of construct validity, by constructing a nomological network. They proposed that the constructs that a test is intended to measure could be represented by a nomological network which included a theoretical framework (for what was being measured) and an empirical framework (for how it was going to be measured). Any associations between the two networks would need to be specified.

Thus, construct validity became the third 'type' of validity in thinking around this time. Construct validity served the purpose of inferring "the degree to which the individual possesses some hypothetical trait or quality (construct) ... that cannot be observed directly" by determining "the degree to which certain explanatory concepts or constructs account

for performance on the test ... through studies that check on the theory underlying the test" (AERA, APA and NCME, 1966, pp.12–13). The 1966 *Standards* distinguished construct validity from other forms of validity in the following way: "Construct validity is ordinarily studied when the tester wishes to increase his understanding of the psychological qualities being measured by the test ... Construct validity is relevant when the tester accepts no existing measure as a definitive criterion" (AERA, APA and NCME, 1966, p.13).

Essentially, construct validity attempted to make a link between assessment performance and pre-conceived theoretical explanations, in other words, to determine the consistency between observed performance on an assessment and its related underlying construct theory. One development of interest at this time came from Campbell and Fiske (1959) who proposed the multi-trait multi-method approach to validation. This included the introduction of two new concepts – convergent validity (the degree to which the test correlates with established tests or assessments purporting to measure similar constructs) and discriminant validity (the degree to which the test does not correlate with measures of different constructs). In practical terms this led to further validation methods involving the use of correlations between different measures, in order to evaluate the likelihood of similar constructs being assessed.

Important features of Cronbach and Meehl's (1955) construct model served as a general methodology for subsequent validation. They emphasised the need for extensive validation efforts, the need for an explicit statement of the proposed interpretation prior to evaluation and the need to challenge proposed interpretations and consider alternate interpretations.

Meehl and Challman (APA, 1954) and Cronbach and Meehl (1955) argued that construct validity offered an alternative to the criterion-based and content-based models. Shortly after the publication of Cronbach and Meehl's (1955) paper, Loevinger (1957) suggested that construct validity was an overriding concern subsuming the content and criterion models. She contended that only construct validity provided a scientifically useful basis for establishing validity. Her assertions foreshadowed Messick's unified view of validity by thirty years reflecting as it did the scientific principles of construct validity.

Kane (2006) asserts that construct validity is deeply based in logical positivistic assumptions which require a coherent and well-articulated theory from which to ground validity claims.

1955–1989: Evolution of the construct validity model

The model of construct validity posited by Cronbach and Meehl appeared to pave the way for validity thinking for the next decade or so, though the model was subject to significant refinement. In 1971, Cronbach wrote the second chapter on validity for *Educational Measurement* thereby adding to and developing Cureton's (1951) position. In his chapter, Cronbach gave construct validity more centrality in relation to the general conception of validity than had the 1966 Standards (AERA, APA and NCME, 1966). Whilst, he continued to maintain the relevance of the triune nature of validity, he likened validity research to the evaluation of a scientific theory as characterised in 'construct validity'. Cronbach argued that most educational assessments involved constructs: "whenever one classifies situations, persons, or responses, he uses constructs" (1971, p.462) and that, "Every time an educator asks 'but what does the instrument really measure?' he is calling for information on construct validity" (1971, p.463).

Cronbach defined validity in terms of interpretations and a range of potential uses and, like his predecessor Cureton, emphasised that validity is not an inherent property of a test but must be evaluated for each testing application:

> *Narrowly considered, validation is the process of examining the accuracy of a specific prediction or inference made from a test score ... More broadly, validation examines the soundness of all interpretations of a test – descriptive and explanatory interpretations as well as situation-bound predictions*. (Cronbach, 1971, p.443)

Within the compass of validity studies Cronbach also included evaluation of decisions and actions based on test scores as well as descriptive interpretations. Cronbach articulated a broad view of validation as involving the evaluation of the interpretations of assessment outcomes and argued that validation focuses on the "accuracy of a specific prediction or inference made from a test score" (1971, p.443). Cronbach (1971) also distinguished a number of approaches to validation, elaborating types of validation needed to support decision-oriented test use. He differentiated validity for selection from validity for placement and emphasised the need to integrate different kinds of validity evidence in evaluating the proposed interpretations and uses of test scores.

Echoing the sentiments expressed in the 1966 *Standards*, the 1974 *Standards* listed four types of validity associated with "four independent kinds of inferential interpretation" (1974, p.26) – predictive and concurrent validities, content validity and construct validity. At this time, the *Standards* explicitly stated validity in terms of its specific intended purpose and context: "No test is valid for all purposes or in all situations or for all groups of individuals" (APA, 1974, p.31).

Unlike criterion-based validation (in which the generation of a correlational index could support validity), or content-based validation, (in which experts attest to the validity of a test's content), construct validation necessitated extensive research effort. Methods employed in construct validation helped determine the link between observed assessment performance and its related construct theory – construct validation being associated with theoretical variables for which "there is no uniquely pertinent criterion to predict, nor is there a domain of content to sample" (Cronbach, 1971, p.462).

Educational measurement theorists throughout this period were beginning to understand that the test itself was not validated; rather, the focus of validation should be the inferences and decisions derived from scores on the test. Alongside this increased awareness was a recognition that multiple measures and multiple evidential sources should be taken into consideration when validating assessment inferences, especially in relation to complex domains.

Towards the end of the 1970s, there existed a tension between major validity theorists who regarded construct validity as dominant model pushing towards a more unified approach to the theory of validity (Cronbach, 1989; Guion; 1977, 1980; Messick, 1975, 1981; Tenopyr, 1977) and those (predominantly assessment users who saw the practical uses of predictive, content, and criterion validity) who continued to work from multiple validity frameworks.

Between the early 1950s and the late 1970s a practice had emerged whereby a 'toolkit' of different models was developed for validating

educational and psychological tests – different models to be employed for different assessments.

The 1980s

By the 1980s the construct model had been adopted as a general approach to validity (Anastasi, 1986; Embretson, 1983; Messick, 1980, 1988, 1989). Messick adopted a broadly defined version of the construct model as a unifying framework for validity. Messick perceived validity as a unified concept and that validity measures are not singular; rather, validity is an ongoing activity that relies on multiple evidence sources. According to Messick (1988, p.35): "from the perspective of validity as a unified concept, all educational and psychological measurement should be construct-referenced because construct interpretation undergirds all score-based inferences – not just those related to interpretive meaningfulness but also the content- and criterion-related inferences specific to applied decisions and actions based on test scores."

In his seminal treatise on validity in the third edition of *Educational Measurement*, Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (1989, p.13). This definition resonates with the definition provided in the most recent version of the *Standards* (AERA, APA and NCME, 1999). Messick (1989) conceptualised validity in terms of value implications and social consequences of testing outcomes. He emphasised validity as an evaluative process focusing on inferences derived from assessment scores (not the assessment itself) and the actions resulting from those inferences. Messick argued that validity extends beyond test score meaning and includes aspects related to score relevance and utility, value implications, and social consequences.[2]

In challenging the 'unholy trinity' of validity, Messick perceived score meaning and construct validity as the underlying objective of all test validation: "validation is a matter of making the most reasonable case to guide both current use of the test and current research to advance understanding of what the test scores mean … To validate an interpretive inference is to ascertain the degree to which multiple lines of evidence are consonant with the inference, while establishing that alternative inferences are less well supported" (1989, p.13).

Messick argued that validation "embraces all of the experimental, statistical, and philosophical means by which hypotheses and scientific theories are evaluated" (1989, p.14) and entails:

- determining "the degree to which multiple lines of evidence are consonant with the inference, while establishing that alternative inferences are less well supported" (1989, p.13).
- "appraisals of the relevance and utility of test scores for particular applied purposes and of the social consequences of using the scores for applied decision making" (1989, p.13).

It is important to stress that validity conceptualised as a unified view did not in any way diminish content or criterion sources of evidence but instead subsumed them in an attempt to build a robust argument for validity. Moreover, the unified approach permitted a fusion of competing theories and validation methodologies. A key point was the idea that a unified, though multi-faceted concept of validity, constituted the foundation for contemporary validity theory.

Contemporary validity theory

Ratcliffe (1983) observed that "quite different notions of what constitutes validity have enjoyed the status of dominant paradigm at different times, in different historical contexts, and under different prevailing modes of thought and epistemology" (p.158). Echoing Ratcliffe's sentiments, Moss, Girard and Haniford (2006) suggest that validity theory can be understood "as an intellectual framework or set of conceptual tools that shapes both our understanding and our actions" (p.109) and as "the representation of an epistemology – a philosophical stance on the nature and justification of knowledge claims – which entails a philosophy of science" (p.110). The epistemological shift in validity theory from a positivistic[3] to a post-positivistic orientation[4] (Moss *et al*., 2006) – described elsewhere by Geisinger (1992) as a 'metamorphosis' – has brought about a variety of epistemological and methodological perspectives within contemporary validity theory (DeLuca, 2009).

In the fourth and latest edition of *Educational Measurement*, Kane (2006) calls for multi-perspective validity arguments to justify test use. Citing House's (1980) logic of evaluation and Cronbach's (1988) earlier work on validation as an evaluation argument, Kane proposes an argument-based approach to validity. Kane's validation framework is congruent with the approach to validation suggested by the current version of *Standards* (AERA, APA and NCME, 1999) and resonates with Messick's 1989 chapter on validity.

Kane (2006) perceives the validation process as the assembly of an extensive argument (or justification) for the claims that are made about an assessment. According to Kane, "to validate a proposed interpretation or use of test scores is to evaluate the rationale for this interpretation or use. The evidence needed for validation necessarily depends on the claims being made. Therefore, validation requires a clear statement of the proposed interpretations and uses" (2006, p.23). Cronbach also conceptualised validity arguments as serving an evaluative function stating that, "validation of test or test use is evaluation" (1988, p.4).

Kane proposed that any validation activity should necessarily entail both an *interpretive* argument (in which a network of inferences and assumptions which lead from scores to decisions is explicated) and a *validity* argument (in which adequate support for each of the inferences and assumptions in the interpretive argument is provided and plausible alternative interpretations are considered).

An argument-based approach to validation is perceived to constitute a compromise between complicated validity theory and a requirement to present a case for the defensibility of using a test for a specified purpose. The force of an argument-based approach to validation is that it:

- ensures that the task of validating inferences is both scientifically sound and logistically manageable;
- provides guidance in apportioning research resource;
- enables estimates of progress in the validation effort to be made;
- and facilitates identification of the various sources of validity evidence that would support or refute the inferences specified on the basis of test scores.

2. A criticism of construct validity as the framework for a unified model of validation was that it did not provide clear guidance for the validation of a test score interpretation or use.
3. The positivistic position assumes a reality that is independent of human perception and therefore draws a distinction between facts and values (Denzin & Lincoln, 2008).
4. The post-positivistic mode of inquiry recognises truth as socially constructed, situational and subjective (Denzin & Lincoln, 2008).

To claim that an interpretation or use of a test is valid is "to claim that the interpretative argument is coherent, that its inferences are reasonable, and that its assumptions are plausible" (Kane, 2006, p.23). In terms of Kane's framework, validation activity requires sufficient evidence that: the test actually measures what it claims to measure; the test scores demonstrate reliability; and that the test scores manifest associations with other variables in a way that is compatible with its predicted properties.

The role of consequences in validity

The most recent version of the *Standards* (AERA, APA and NCME, 1999) identifies five sources of validity evidence, one of which is "evidence based on consequences of testing" (1999, p.16).[5] Describing consequences, the *Standards* "distinguish between evidence that is directly relevant to validity and evidence that may inform decisions about social policy that falls outside the realm of validity" (1999, p.16). That the role of consequences should be included in the *Standards* as a potential source of validity evidence is undoubtedly a result of Messick's (1989) hugely influential chapter in which he formalises the consequential bases of test interpretation and test use. Messick's (1989) integration of both evidential and consequential sources of evidence have served to appreciably widen the compass of validity inquiry by including social and value-laden aspects of assessments thereby extending traditional measurement boundaries into issues relating to policy – what Kane (2001) has termed the prescriptive part of a validity argument. This has necessitated the requirement for evidence about the social consequences of test use (Cronbach, 1988; Messick, 1989, 1994; Shepard, 1993; Linn, 1997). However, whether Messick's definition of validity included evidence about all consequences of assessment as validity is fiercely contested. Even Shepard, an advocate of consequential validity, acknowledges that ''there is a great deal more in what Cronbach and Messick have suggested than is acknowledged or accepted by the field'' (1993, p. 406).

The role of consequences in testing has become a highly controversial issue within contemporary validity debate (Crocker, 1997; Brennan, 2006). Brennan states, "since it is now almost universally agreed that validity has to do with the proposed interpretations and uses of test scores, it necessarily follows that consequences are a part of validity" (2006, p. 8). However, there is considerable disagreement regarding the role that the consequences of test score use plays in validity theory. The importance of the debate is most clearly illustrated by the fact that two entire issues of the journal *Educational Measurement: Issues and Practice* were given over to such concerns in 1997 and 1998.

Of course the role of consequences in testing is not new. Cureton (1951) acknowledged consequences as being a part of validity in his chapter and Kane (2006) maintains that consequences have always played an integral role in validation. There exists within the educational measurement community, therefore, general agreement that evaluating consequences is important. What is contentious, however, is the validation of both intended consequences (claimed outcomes) and unintended or negative consequences of test use. Since consequences reflect the effects or impacts of test usage, evaluating intended consequences is ostensibly an attempt to evaluate the extent to which a test fulfills its specified purpose or proposed use. For a full evaluative treatment of all consequences to be complete, analysis of evidence would require monumental validation effort especially if it is to include an exploration of unintended consequences.

Some measurement theorists (Maguire, Hattie and Haig, 1994; Crocker, 1997; Green, 1998; Mehrens, 1997; Popham, 1997; Borsboom, Mellenbergh and van Heerden, 2004) have argued for a limited and more technical definition of validity that emphasises the descriptive interpretation of scores. Whilst they suggest that consequences are crucial to social research they nevertheless categorise them as being outside validity theory. According to Maguire *et al*., "Consequences should be moved out from the umbrella of construct validity and into the arena of informed social debate and formulated into ethical guidelines" (1994, p.115). Others, however, embrace a broader view of validity arguing that assessments should be contextualised within their consequential outcomes (e.g. Linn, 1997; Messick, 1989; Moss, 1998; Shepard, 1997; Kane, 2001).

Summary

Within the sphere of educational assessment there is now broad agreement regarding Messick's (1989) definition of validity as about the appropriateness of the inferences and uses of assessment outcomes (though this is by no means universal, see for example, Borsboom, 2006; Borsboom, Mellenbergh and van Heerden, 2004; Lissitz and Samuelsen, 2007). Validition is perceived by Kane (2006) to be a judgement of the degree to which arguments support those proposed interpretations and uses. Following an extensive review of the literature, Sireci (2007, 2009) summarises the fundamental features of validity in the following ways (2007, p.477):

- validity is not an inherent property of a test but refers to the specified uses of a test for a particular purpose;
- validity refers to the proposed interpretations or actions that are made on the basis of test scores;
- in order to evaluate both the usefulness and appropriateness of a test for a particular purpose multiple sources of evidence are required;
- sufficient evidence must be collected to defend the use of the test for a particular intended purpose;
- the evaluation of validity is neither static nor a one-time event but a continuing process.

Messick (1989) argued that "validity is an evolving property and validation is a continuing process" (p.13). The contemporary conceptualisation of validity cannot be considered definitive, but as the current most accepted notion. This, and particularly the role of consequences as part of validity, is likely to continue to evolve over time.

5. The other sources of validity evidence include *test content*; *response processes*; *internal structure*; and *relations to other variables*

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1954/1955). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques [supplement]. *Psychological Bulletin*, **51**, 2, Pt.2, 201–238.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, **37**, 1–15.

Binet, A. (1905). New methods for the diagnosis of the intellectual level of subnormals. *L'Année Psychologique*, **12**, 191–244.

Binet, A., & Henri, B. (1899). La psychologic individuelle. *Amiee Psychol*., **2**, 411–465.

Bingham, W. V. (1937). *Aptitudes and aptitude testing*. New York: Harper.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, **71**, 425–440.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, **111**, 1061–1071.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In: R. L. Brennan (Ed.), *Educational Measurement*, (4th ed.), 1–16. Westport, CT: Praeger.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, **56**, 81–105.

Crocker, L. (1997). The great validity debate. *Educational Measurement: Issues and Practice*, **16**, 2, 4.

Cronbach, L. J. (1971). Test validation. In: R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.), 443–507. Washington, DC: American Council on Education.

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

Cronbach, L. J. (1989). Construct validation after thirty years. In: R. L. Linn (Ed.), *Intelligence: Measurement, theory and public policy*, 147–171. Urbana: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, **52**, 281–302.

Cureton, E. E. (1951). Validity. In: E. F. Lindquist (Ed.), *Educational measurement*, 621–694. Washington, DC: American Council on Education.

DeLuca, C. (2009). *Contemporary Validity Theory in Educational Assessment: Integrating an Interpretivistic Approach through Case Study Methodology*. Paper presented at the annual meeting of the Canadian Society for the Study of Education, Ottawa, Ontario, May 2009.

Denzin, N. K., & Lincoln, Y. S. (2008). *Strategies of Qualitative Inquiry*. (3rd ed.). Thousand Oaks, CA: Sage.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, **93**, 179–197.

Geisinger, K. F. (Ed.) (1992). *The psychological testing of Hispanics*. Washington, DC: APA.

Goodwin, L. D. (2002). Changing conceptions of measurement validity: An update on the new standards. *Journal of Nursing Education*, **41**, 100–106.

Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practice*, **17**, 2, 16–19, 34.

Guion, R. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement*, **1**, 1–10.

Guion, R. M. (1980). On trinitarian conceptions of validity. *Professional Psychology*, **11**, 385–398.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, **6**, 427–439.

Gulliksen, H. (1950). Intrinsic validity. *American Psychologist*, **5**, 511–517.

House, E. T. (1980). *Evaluating with validity*. Beverly Hills, CA: Sage Publications.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, **38**, 4, 319–342.

Kane, M. T. (2006). Validation. In: R. L. Brennan (Ed.), *Educational Measurement*. (4th ed.). 17–64. Westport, CT: American Council on Education/Praeger.

Kelley, T. L. (1927). *Interpretation of Educational Measurements*. New York: New World Book Company.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, **16**, 2, 14–16.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, **36**, 437–448.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, **3** (Monograph Supplement 9), 635–694.

Maguire, T., Hattie, J., & Haig, B. (1994). Construct validity and achievement assessment. *Alberta Journal of Educational Research*, **40**, 109–126.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, **16**, 2, 16–18.

Messick, S. (1975). The standard program: Meaning and values in measurement and evaluation. *American Psychologist*, **30**, 955–966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, **35**, 1012–1027.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher*, **10**, 9–20.

Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In: H. Wainer and H. Braun (Eds.), *Test validity*, 33–45. Hillsdale, NJ: Lawrence Erlbaum.

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher,* **23**, 2, 13–23.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, **17**, 2, 6–12.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education*, **30**, 109–162.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society A*, **187**, 253–318.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice*, **16**, 2, 9–13.

Ratcliffe, J. W. (1983). Notions of validity in qualitative research methodology. *Knowledge: Creation, Diffusion, Utilization*, **5**, 147–167.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, **19**, 405–450.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, **16**, 2, 5–8, 13, 24.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, **36**, 477–481.

Sireci, S. G. (2009). Packing and Unpacking Sources of Validity Evidence. In: Lissitz, R. W. (Ed.), *The Concept of Validity: Revisions, New Directions, and Applications*. IAP: Charlotte, NC.

Spearman, C. (1904). General intelligence: objectively determined and measured. *American Journal of Psychology*, **15**, 201–293.

Suppe, P. (1977). *The structure of scientific theories*. Urbana, IL: University of Illinois Press.

Tenopyr, M. L. (1977). Content-construct confusion. *Personnel Psychology*, **30**, 47–54.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurements of educational products. *National Society for the Study of Educational Products: Seventeenth Yearbook*. 16–24.

Does doing Critical Thinking AS level confer any advantage for candidates in their performance on other A levels?

Beth Black and Tim Gill Research Division

Critical Thinking can be defined as analytical thinking which underlies all rational discourse and enquiry (Black *et al*., 2008). It is of some interest whether when taught as a separate course, it can be transferred by students to other subject domains and improve their performance in them. In the UK context, Critical Thinking AS level was introduced in schools in 2001 and, as such, represents the catalyst for a large scale introduction of this discipline into schools.

There is now much research that shows that the teaching of Critical Thinking (CT) does indeed improve critical thinking skills. Abrami *et al*. (2008) provides an excellent meta-analysis of studies into the effectiveness of teaching CT. The average effect size was 0.34, indicating that CT interventions tend to have a small to moderate impact upon the development or enhancement of CT skills and dispositions. In one of Abrami *et al*.'s sub-analyses, the 117 studies included in the meta-analysis are divided into one of four types based on the instructional method of the intervention – general, infusion, immersion and mixed[1]. In a 'general' course, CT is taught without any other specific subject matter or domain content – in other words, the main (and only) objectives of the course are to improve CT skills and dispositions. For both 'infusion' and 'immersion' courses, CT is delivered through other subject content, though where they differ is that CT principles and learning objectives are explicit in an 'infusion' approach, while implicit in an 'immersion' approach. Finally, the 'mixed' approach again involves teaching CT through another subject, though it is delivered as an independent track within that subject. The meta-analysis revealed that there were positive effect sizes for all types of intervention. However, immersion (with no explicit CT objectives) was least effective (effect size = 0.09); while the mixed approach was the most effective (effect size = 0.94) with the general and infusion approaches also having moderate to large effect sizes (0.38 and 0.54 respectively).

The result for the 'general' approach is quite interesting given John McPeck's (1981) well-known objections to CT being taught in such a way, as a standalone subject. His point is that one always has to think about something.

> *In isolation from a particular subject, the phrase "Critical Thinking" neither refers to nor denotes any particular skill. It follows from this that it makes no sense to talk about Critical Thinking as a distinct subject and that it therefore cannot be profitably taught as such. To the extent that critical thinking is not about a specific subject X, it is both conceptually and practically empty.*

Thus, Abrami *et al*.'s research appears to contradict this view and show that CT, taught generally as a standalone subject, can improve CT skills.

1. This classification is based upon Ennis's (1989) typology.

However, there is less research which shows whether CT skills when taught (in any of the four approaches described above) can be profitably transferred to other subject domains. This is of keen interest since much of the rhetoric around CT as a worthwhile educational goal rests on the notion that it is not just good *in itself* but "being able to think critically is a necessary condition of being educated in a more general sense" (Norris, 1985). Again, there is much speculation as to the best way to deliver CT so as to foster transferable CT skills and dispositions (e.g. Brown, 1997; Halpern, 1998).

In the UK context, from a survey of CT teachers (Black, 2010), we know that CT tends to be delivered separately or discretely – as the 'general' approach, rather than within other subjects. This survey also revealed that the overwhelming majority of respondents (95.7% of all respondents) believed that students did (spontaneously) transfer these skills to other subjects to the benefit of their performance, skills and understanding in other subjects. Of course, the crucial word here is 'believed'. It was the belief or perception of teachers, based on their own (anecdotal) experiences with their students, rather than hard evidence:

> *…the majority [of students] find it quite useful and they now write better essays or think more logically. One said "it has changed my whole way of thinking".*

As well as based upon their understanding of how these skills form a fundamental part of other educational endeavour:

> *[CT] complements analytical requirement in many subjects…Many of our "most-improved" students in year 13 took CT… perhaps due to developing transferable skills.*
>
> *Many subjects call for reasoned arguments. What better way to prepare them?*

Therefore, we were particularly interested to see whether there was any data to support these views that students who have taken CT do better in their other subjects.

This report looks at the performance at A level of candidates who had taken CT AS level, in comparison to candidates who had not taken the CT AS level. It was hypothesised that CT skills are transferable and can be applied to other subjects in a beneficial way. Thus candidates gaining good CT skills at AS level may improve their performance at A level.

The hypothesis that we put forward here is that candidates who took CT, and gained a good grade in it performed better in their A levels than similar candidates who did not take CT. If this is shown to be the case then we can infer that the skills gained by taking the CT AS level were beneficial to the candidates in their other A levels. Of course we cannot prove this association, because many other factors influence how well candidates perform in their A levels.

Data and methods

Data taken from the NPD databases for 2005 and 2006 were used for this research. These are databases of all exams taken in England and Wales by pupils of different ages. From these it was possible to identify candidates taking CT AS level, and follow them through to their A level results.

The first analysis looked at whether candidates who had performed well in the CT AS level (A and B grade candidates) performed better at A level on average than candidates who did not take CT at all. Candidates getting grades A or B at CT AS level in 2005 were identified in the database. These candidates were then matched to a set of candidates not taking CT by ability (as measured by GCSE mean grade) and the A level results in 2006 for the two different groups were compared.

To choose the matched candidates, the mean GCSE was calculated by converting grades into numbers, with 8 for an A*, 7 for an A and so on down to 0 for U. The distribution of GCSE mean grade for the candidates receiving grades A and B for CT AS level was inspected (n=2208). These pupils were divided up into 20 approximately equal groups (in terms of numbers) by mean GCSE grade. A random sample of candidates was then taken, matched to each of the 20 groups, from the remaining candidates in the database (the non-CT group). For example, the bottom group of CT candidates consisted of 108 pupils with a mean GCSE of between 3.86 and 5.86. A matching sample of 108 was (randomly) taken from the group of non-CT candidates with a GCSE mean grade of between 3.86 and 5.86.

This was done for each of the 20 groups. The 20 random samples were joined together to create one overall matching dataset. The following summary statistics demonstrate that the CT and non-CT groups were well matched:

Table 1: Summary statistics for matched groups

	N	*GCSE Mean*	*SD*	*Min*	*Max*
CT	2208	7.18	0.67	3.86	8.00
Non-CT	2208	7.17	0.69	4.00	8.00

A further analysis was undertaken, comparing the A level performance of all the CT candidates with the performance of all candidates taking A levels in 2006.

Results

Comparison of means

First we looked at the mean A level grades and total A level score for the two groups of candidates (A or B grade CT candidates and non-CT). To calculate the means and totals each grade was transformed into a number, with 10 for a grade A, 8 for a grade B and so on, down to 0 for a grade U. A statistical test (Kolmogorov-Smirnov or K-S test) was used to determine if the difference between the groups in the distribution of their mean or total A level scores was statistically significant or could be attributed to chance[2]. The results are shown in Table 2.

The mean and total A level scores were clearly higher for the high performing CT candidates on average, compared to the non-CT candidates. The difference in the mean A level (9.12–8.68 = 0.44)

2. It was not possible to test for difference in the means using a t-test as the distributions of mean and total grades were not normal.

Table 2: Overall mean A level performance for CT and non-CT candidates

	Group	*N*	*Mean*	*Std. Deviation*	*Sig of K-S Test*
Mean A level	Non-CT	7295	8.68	1.63	<0.001
	CT	7691	9.12	1.20	
Total A level	Non-CT	7295	32.04	9.40	<0.001
	CT	7691	34.39	8.89	

translates to around a quarter of a grade per A level. The effect is not very large, but would be the equivalent of a grade for a candidate taking four A levels.

According to the K-S test the differences in the distributions of both the mean and total A level scores were highly significant. The figure in the final column gives the probability of a difference the same as or larger than observed occurring if there was actually no difference between the two groups. A figure of less than 0.05 is generally considered to be significant, and less than 0.01 highly significant.

The direction of the difference can be seen by sketching the cumulative distributions functions of A level grade and total A level score for the two groups:

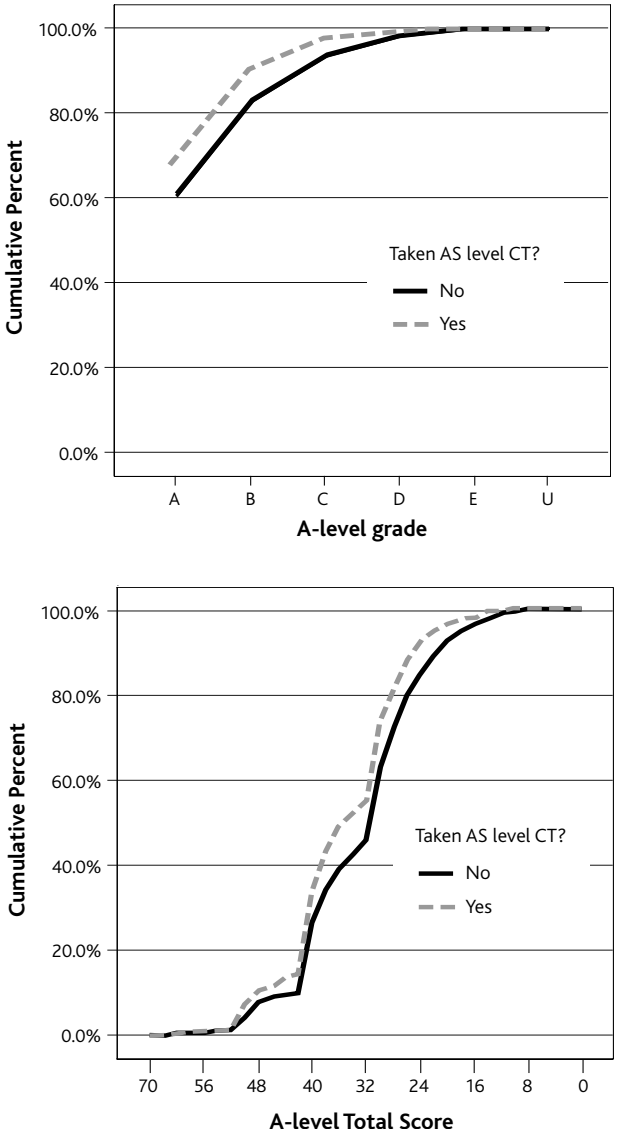


Figure 1: Cumulative frequency distributions of A level grade and total A level score

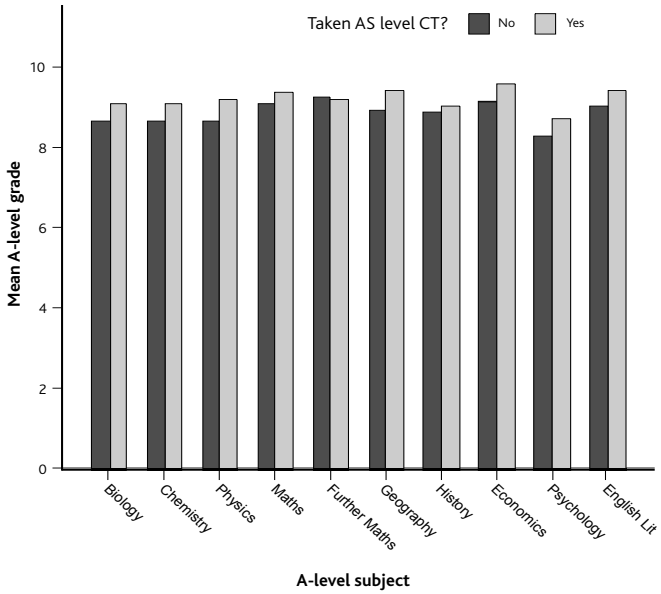


Figure 2 : Mean A level grades for individual subjects

It is clear that the distribution for the CT candidates is further to the left, meaning a larger percentage of this group were towards the top in terms of A level grade than the non-CT students. For instance, in the left hand figure for the non-CT students just over 80% of the grades received were at least a grade B, compared to about 90% of the grades received by the CT students.

We also investigated performance at A level in the most popular subjects individually. For this we selected the candidates from the groups (CT and non-CT) taking each individual subject. This meant that for some of the subjects the candidates in the two groups were no longer exactly matched for ability. Thus some caution should be exercised with the results for these subjects. The mean A level grade for each group in each subject is shown in Figure 2. This data is also displayed in Table 3, along with the number of students and the outcome of the K-S Test. To assist interpretation the GCSE mean grades for candidates in each group taking the subject in question are also listed in the table:

Table 3: Mean A level grades for individual subjects

	GCSE mean grade		*A level candidates*		*A level mean grade*		*Sig of K-S Test*
	Non-CT	*CT*	*Non-CT*	*CT*	*Non-CT*	*CT*	
Biology	7.34	7.33	670	589	8.70	9.10	0.010
Chemistry	7.40	7.44	658	559	8.68	9.14	0.044
Physics	7.37	7.40	364	359	8.70	9.22	0.018
Maths	7.41	7.40	846	801	9.12	9.38	0.065
Further Maths	7.54	7.44	156	182	9.28	9.22	1.000
Geography	7.17	7.21	296	236	8.95	9.44	0.059
History	7.27	7.21	448	681	8.88	9.09	0.559
Economics	7.36	7.44	174	264	9.17	9.60	0.037
Psychology	6.86	6.79	254	267	8.31	8.73	0.333
English Lit	7.25	7.29	505	676	9.09	9.45	0.002

In all the subjects apart from Further Maths the group performing well in CT had a higher mean A level grade in the subject than the non-CT. The K-S test shows there was a significant difference in the distribution of A level grade between groups in several of the subjects. However, we must also consider any differences in the mean GCSE grades.

For Biology and Maths there was virtually no difference between the GCSE mean grades of the two groups, so we can assume they are matched. The K-S test was significant for Biology, so we have evidence that the CT candidates performed better in this subject. However, the test was not significant for Maths, so there was no evidence of improved performance.

For Chemistry, Physics, Economics and English Literature the CT candidates had slightly higher GCSE mean grades, so the significantly better performance of this group at A level was not as high as suggested in the table, and would potentially be non-significant if we had data that matched exactly on prior attainment. However, in each case the difference in GCSE mean grade was small in comparison to the differences in mean A level grade so it is still probable that a significant difference was present.

For Psychology and History, although the difference between the two groups at A level was not significant, the non-CT group had a higher GCSE mean grade. Thus it may be that if the groups were matched more exactly, the performance of the CT group would have been significantly better at A level.

Finally, there was no significant difference between the two groups in their A level Geography performance and, as the CT candidates had a slightly higher GCSE mean grade, there was certainly no evidence that these candidates performed better at A level.

In summary, there is evidence that candidates who achieved high grades in AS level CT performed better overall at A level than candidates who did not study CT at all. There is evidence that this advantage presents itself across a wide range of subjects, in sciences, social sciences and arts subjects. This backs up the hypothesis that CT skills are transferable and applicable to a wide range of subjects.

Regression analysis

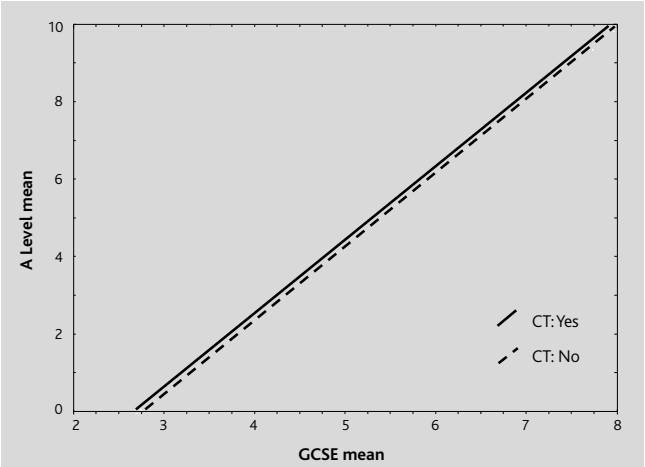
In the previous section we only selected candidates who received a grade A or B in Critical Thinking at AS level. An alternative way of analysing the data is to undertake a linear regression on overall A level performance for *all* candidates. This predicts a mean A level score (and separately a total A level score), based on certain variables. We allowed for previous attainment by including candidates' GCSE mean grade in the model. A variable indicating whether or not the candidate studied AS level CT was also included, which enabled the impact of taking this qualification to be analysed, for a given level of prior attainment. It was also possible to analyse the impact of having received a particular grade on the CT AS level.

Mean A level grade

Figure 3 shows some basic regression output from a model with mean A level grade as the dependent variable and GCSE mean grade and whether or not the candidate had studied CT as the predictor variables.

The R square is a measure of the amount of variation in the dependent variable that can be accounted for by variations in the predictor variables. Thus, 52% of the variation is explained by the regression model, which is reasonable.

We can see from the variables table that both the predictor variables are highly significant (Sig < 0.01). This means we have evidence that changes in these are associated with changes in the mean A level grade. This effect is quantified by B, which is the change in the dependent variable as a result of a unit increase in the predictor variables. Thus the



Variable	B	Sig	R	R Square
Constant	-5.34	<0.01	0.72	0.52
GCSE mean grade	1.92	<0.01		
Critical Thinking	0.20	<0.01		

Figure 3: Regression output on mean A level

model predicts that an increase in GCSE mean grade of 1 unit (equivalent to 1 grade) leads to an increase in A level mean grade of 1.92 (equivalent to just under one grade).

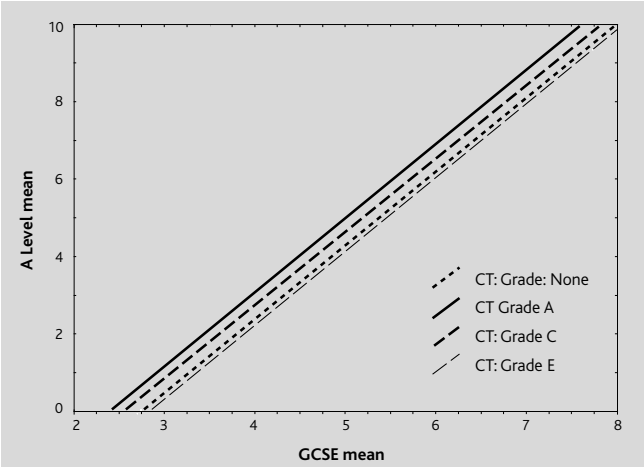
The CT variable is specified as a 1 if the candidate has taken the AS level and a zero if not. Thus, according to the model, having taken the AS level increases (on average) candidates mean A level grade by 0.20, or about one tenth of a grade.

The graph in Figure 3 can help with interpreting the model. This plots GCSE mean against A level mean (as predicted by the model) for the CT and non-CT groups. This demonstrates that for a particular level of GCSE mean, the model predicts a higher A level mean grade for candidates who took the CT AS level, than for those who did not. However, the difference is clearly not very large.

The second model, which is shown in Figure 4, also took into account the grade received by the candidates who took the AS level in CT.

The R square is very similar to the previous model. Once again all of the predictor variables are highly significant. The interpretation of this model is, however, more complicated. The grade received at AS level has been split up into a set of 'dummy' variables, one for each grade (apart from U). The coefficients in the table (B) represent the difference in the predicted mean A level for candidates who have received the particular grade in CT *in comparison to a candidate who received a grade U*. So, a candidate receiving an A grade has a predicted mean A level grade 0.91 higher than a U grade candidate.

To compare the performance of a candidate getting a particular grade on the CT AS level with one not taking the qualification at all, a combination of the coefficient for the grade and the CT coefficient is required. For example, imagine two candidates with the same GCSE mean grade, one having taken the AS level in CT and received a grade C, and the other having not taken CT. The predicted mean A level grade for the candidate who took the CT AS level will be 0.69–0.27 = 0.41 higher than the candidate not taking CT. Thus the overall effect is an increase in predicted mean A level grade of 0.41, or around one fifth of a grade. For a candidate with a grade B the overall predicted increase is 0.67 (0.94–0.27) and for a grade A candidate it is 0.64 (0.91–0.27), both of which are about one third of a grade. For a candidate who took three A levels this amounts to around one grade overall.



Variable	B	Sig	R	R Square
Constant	-5.25	<0.01	0.72	0.52
GCSE mean grade	1.91	<0.01		
Critical Thinking	-0.27	<0.01		
Grade A	0.91	<0.01		
Grade B	0.94	<0.01		
Grade C	0.69	<0.01		
Grade D	0.51	<0.01		
Grade E	0.22	<0.01		

Figure 4: Regression output on mean A level, including grades achieved on CT

Again, a graph can aid interpretation of this result. Figure 3 plots the GCSE mean against predicted A level mean for candidates achieving different grades in their AS level CT, and for the non-CT group[3]. Thus a grade E candidate had a very slightly lower predicted mean A level grade than a non-CT candidate with the same prior attainment. Candidates getting a grade A or C had a higher predicted mean A level than a non-CT candidate with the same prior attainment.

It is worth noting the unexpected result that the coefficient for a grade B in CT is higher than that for a grade A. Inspection of the distribution of mean A level showed this to be a 'ceiling' effect. Of the candidates who received an A or a B at CT a large proportion (1,414 out of 3,357) received all grade As at A level, giving them the maximum mean A level score of 10, and many others had a mean A level grade of 9 or more. Thus the level of discrimination was not enough to be able to distinguish between the CT grade A and grade B candidates.

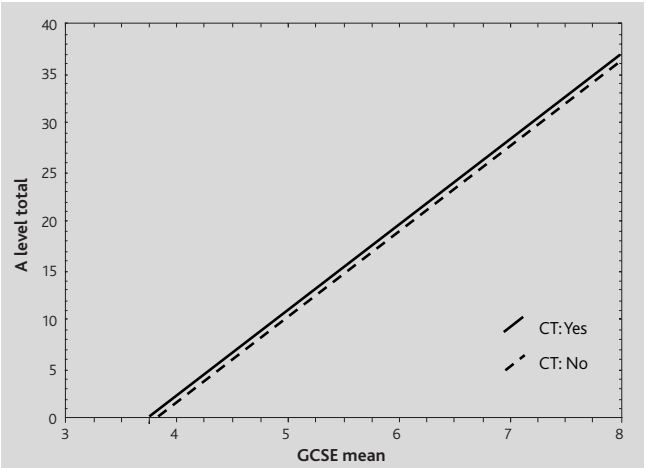
Total A level grade

We repeated both models using total A level score as the dependent variable. Figures 5 and 6 have the output from the two models with the same predictor variables as above.

Both models had reasonable R square values and all coefficients were significant. For the overall model, having taken CT increased the predicted total A level score by 0.64, or about one third of a grade. In terms of the individual grades, getting a grade U reduced the predicted total by 1.69, compared to not taking CT, and a grade E reduced it by 0.58. For all other grades the predicted score increased compared to not taking CT, in ascending order of grade. Having a grade A increased it by 3.71, equivalent to almost two grades.

Note that in this case, the grade A coefficient was larger than the grade B coefficient, so as expected getting a grade A gave more of an advantage in terms of total A level score than getting a grade B. This was because

3. The lines for grade A, C and E only are shown to avoid crowding the graph too much.



Variable	B	Sig	R	R Square
Constant	-32.97	<0.01	0.75	0.56
GCSE mean grade	8.66	<0.01		
Critical Thinking	0.64	<0.01		

Figure 5: Regression output on total A level score

there is more discrimination at the top end with regards to total A level score, since the candidates with a mean A level score of 10 (all grade A) are split into those with a total A level score of 30, 40, 50 and 60.

Thus we have further evidence that candidates taking the AS level CT, and getting a reasonably good grade, perform better overall on their A levels the following year. According to the model presented here the improvement for the top candidates is around one third of a grade on the mean A level and around two grades when looking at total A level score.

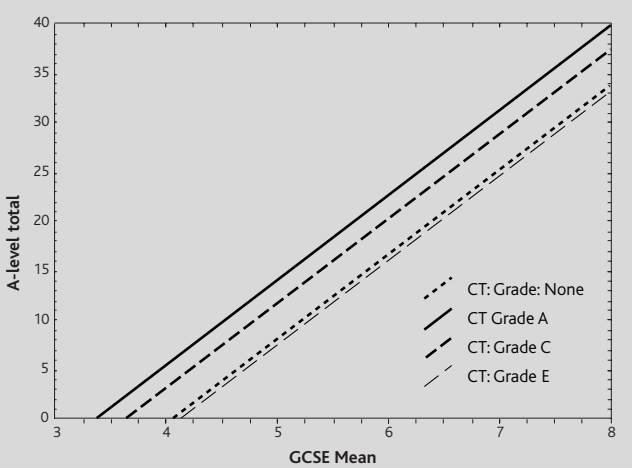
Discussion

We should note some caveats of this research. Although we have shown an association between taking CT at AS level and performing well at A level, we cannot be sure that the former *causes* the latter. It may be that candidates who perform well on CT do so because they already possess the skills and attributes to perform well academically more generally (although this had not differentially benefitted them at GCSE level).

Secondly, since not all schools offer CT, there may be a school effect that we have not been able to identify. For instance, perhaps only the better schools offer it, in which case the candidates in these schools are likely to perform better overall. An alternative analysis would be to use data over time, and see if centres that started teaching CT saw an improvement in the progress of pupils from GCSE to A level in subsequent years, whilst similar centres that did not teach CT improved less or not at all.

However, if we accept the interpretation that studying CT AS level can improve performance in other subjects, it is worth reflecting on this a little further. Piecing these findings together with the survey data (Black, 2010), we might be surprised by any discernible transfer effect for a number of reasons:

- Teachers often reported little or no training in improving their own CT skills or how to teach the discipline.
- Limited resourcing of the courses in terms of amount of dedicated timetabling as well as other resources (e.g. teaching materials).



Variable	B	Sig	R	R Square
Constant	-32.51	<0.01	0.72	0.52
GCSE mean grade	8.59	<0.01		
Critical Thinking	-1.69	<0.01		
Grade A	5.39	<0.01		
Grade B	4.69	<0.01		
Grade C	3.14	<0.01		
Grade D	2.21	<0.01		
Grade E	1.12	<0.01		

Figure 6: Regression output on mean A level, including grades achieved on CT

- A significant minority of centres reported low motivation of staff and students where the CT course was obligatory rather than optional.
- Teachers tended to report that their overall agenda or aim was for students to achieve a good grade in the CT exam, rather than to foster transferable skills and dispositions.

Therefore, if this study does indeed present evidence for the transferability of CT, it might almost be viewed as an unintended (though serendipitous) consequence of delivering the CT AS level.

This research also suggests that it would be of some interest to investigate the mechanisms by which transferability is best fostered *within* this general or standalone approach to teaching.

References

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

Black, B. (2010). "It's not like teaching other subjects" – the challenges of introducing Critical Thinking AS level in England. *Research Matters: A Cambridge Assessment Publication*, **10**, 2–8.

Black, B., Chislett, J., Thomson, A., Thwaites, G., & Thwaites, J. (2008). A definition and taxonomy for Critical Thinking. *Research Matters: A Cambridge Assessment Publication*, **6**, 30–36.

Brown, A. (1997). Transforming schools into communities of thinking and learning about serious matters. *American Psychologist*, **52**, 399–413.

Ennis, R. H. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher*, **18**, 3, 4–10.

Halpern, D.F. (1998). Teaching critical thinking for transfer across domains. Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist*, **53**, 4, 449–55.

McPeck, J. (1981). *Critical thinking and education*. Oxford, UK: Martin Robertson.

Norris, S.P. (1985). Synthesis of Research on Critical Thinking. *Educational Leadership*, **52**, 8, 4.

Comparing the demand of syllabus content in the context of vocational qualifications: literature, theory and method

Nadežda Novaković and Jackie Greatorex Research Division

This article is based on a presentation given at The Journal of Vocational Education and Training 8th International Conference held in Worcester College Oxford, UK, in July 2009. The paper was written at the beginning of a wider research project, conducted within the Research Division. The aim of the wider project is to develop a research instrument for comparing the syllabus demands of cognate units from different types of qualifications. The specific aims of the present article are to review the theoretical approaches, methods and research instruments used to compare vocational qualifications (VQs) in England, with the view to gauging their appropriateness for comparing the demands of different types of qualifications. The wider project is still work in progress.

Abstract

Our literature review considers the methods used in studies comparing the demands of vocational syllabus content in England. Generally, categories of demands are either derived from subject experts' views or devised by researchers. Subsequently, subject experts rate each syllabus on each demand category and comparisons can be made. However, problems with the methods include:

- Some studies over-focus on the cognitive domain rather than the affective, interpersonal and psychomotor domains.
- Experts vary in their interpretations of rating scales.

Therefore, we suggest creating a framework of demands which includes all four domains, based on a variety of subject experts' views of demands. The subject experts might rank each syllabus on each type of demand, thus avoiding the problem(s) of rating scales, and facilitating comparisons between syllabuses.

Introduction

Comparability is a complex area of research and investigation, which has been very prominent in the debate about the quality of summative[1] assessment in England in the past decade. This activity has been fuelled by public expectation that assessment standards should remain constant over time, across subjects, between awarding bodies, between test and task demands and so on.

We were tasked with considering methods for comparing the demands of cognate qualifications including vocational qualifications (VQs)[2] in a situation where performance data and performance evidence were lacking and there was limited access to the assessment tasks. This would result in small and/or unrepresentative samples of performance data, performance evidence and assessment tasks. Given the complexity of comparability research we focus on one aspect of comparability: the demands of different qualifications' syllabus content. There are various definitions of 'syllabus', see Nunan (1988) for a detailed discussion. Here syllabus refers to: the statement of the aims/objectives/purpose of the qualification; what knowledge and skills can be in the summative assessment(s); how this will be assessed; and descriptions of levels of quality of performance (e.g. pass or particular grades).

This article presents a review of the relevant research literature relating to comparability within, or at least partly covering, the context of VQs. Different theoretical approaches, methods and research instruments are discussed with the view to gauging their appropriateness for comparing the demands of different types of qualifications.

Comparability of vocational qualifications

A recent publication on the comparability of assessment standards (Newton *et al*., 2007), contains an appendix of 154 comparability reports. However, only seven of these include a VQ, the remainder relate to general qualifications (GQs)[3], illustrating the disparity between comparability research in VQs and GQs.

This disparity is unsurprising, as researching the comparability of VQs is beset by issues not present in the context of GQs. Johnson (2008) indicates that VQs have lower *assessment density* than GQs, *assessment density* refers to the frequency with which assessors judge the same type of performance evidence in similar contexts. Unlike GCSEs[4] and GCE A levels[5], which are mostly assessed through large-scale examinations, VQs tend to be individualised and partly or wholly assessed by criterion-referenced, outcome-based assessment. Therefore, VQ assessors tend to be assessing each candidate's skills and competence based on the evidence of how they perform on specific

1. Summative assessment is generally used to provide an overall grade or level of achievement for a particular learning programme. Normally summative assessment is used for the purpose of determining who will be awarded a qualification.
2. Vocational qualifications are designed to focus on learning practical skills (OCR, 2009). Vocationally-related qualifications, give a broad introduction to a particular sector, for example the media or health. For the purpose of brevity we use *vocational qualifications (VQs)* to refer to vocational and vocationally related qualifications.
3. General qualifications always include examinations as part of the summative assessment. They are broad in nature rather than focused on any particular work-related area (OCR, 2009).
4. General Certificate of Secondary Education (GCSE). These qualifications are generally taken by 16 year olds at the end of compulsory schooling. Usually students take GCSEs in a series of school subjects. GCSEs are general qualifications.
5. General Certificate of Education Advanced level qualifications. Generally they are divided into an AS qualification, most often taken by 17 year olds, and A2 assessments, mostly taken by 18 year olds. Combined together the results of the AS and A2 assessments give A level results. A levels are general qualifications.

tasks in specific settings. Candidates' skills may be assessed by a broad range of assessments which may vary considerably from one centre to another (for example, the choice between simulated and authentic activities to assess the same skill).

Most comparability studies focussing on a VQ compare it with a GQ. Greatorex (2001) argues that such comparisons might not be robust due to differences between the qualifications which cannot be accounted for by experimental or statistical controls. For instance: different purposes, learner populations, modes of assessment (e.g. examinations, portfolios) and approaches to applying assessment criteria (e.g. compensation, hurdles).

All the above-mentioned factors might have contributed to the relative paucity of research into the comparability of VQs. However, researching issues relating to comparability in VQs is important for several reasons. First, such investigations are likely to help ensure that VQs are perceived as robust qualifications with consistent standards. Some studies have already been carried out to this effect. For instance, Arlett (2002, 2003) conducted two studies comparing the performance standards and demands of VQs across different awarding bodies in the context of VCE[6] Health and Social Care, a new qualification at the time, and found few large differences. However, Arlett (2003) found a perceived difference in the demand of questions. Guthrie (2003) carried out a similar study comparing GCE Business studies and VCE Business. In many ways the examination and syllabus demands of the VCE versus the GCE were found to be similar. However, the differences in demand between the different types of qualifications were:

- GCE syllabuses encouraged a more synoptic approach than the VCE syllabuses.
- VCE syllabuses encouraged the acquisition of Business skills much more than the GCE syllabuses.
- GCE timed examinations were considered more demanding than the VCE timed examinations.

The research into the comparability between GQs and VQs should also go some way to addressing the dilemmas experienced by employers faced by candidates in possession of different awards – are such qualifications of the same standard, what is the common standard that they share, and what exactly are the differences between them? This is also important as some VQs offer an alternative route to higher education. If two entrants for the same university course both fulfil the requirements for gaining a place but one is in possession of A level qualifications while the other has a VQ, the expectation is that these qualifications should share the same standard.

According to McEwen *et al*. (2001) the traditional view of academic qualifications is that they promote deep conceptual understanding, but may lead to superficial understanding, regurgitation for assessment, and knowledge which cannot be applied outside the narrow range of contexts. On the other hand, GNVQs[7] aimed to integrate 'knowing what' and 'knowing how', but students may not be sufficiently exposed to a wide range of conceptual enquiry and cognitive skills might be neglected (McEwen *et al*., 2001). This is linked to the view that VQs are often seen as an 'easier option' to A levels for lower ability students. According to Barry (1997, p. 44) GNVQs received a lot of "bad press" through some academics condoning the GNVQs as a "second rate" alternative to A levels, and suggesting that the skills learners developed during a GNVQ are gone within a year leaving the learners ill-equipped to study for a single honours degree. Defining the shared standard and clearly stating differences between GQs and VQs might help to address some of the preconceptions currently surrounding VQs.

However, it is unclear how the equivalence of standards should be investigated in such complex cases involving assessments of different nature, designed for different populations of students. Pollitt *et al*. (2007) suggest that no definition of comparability should necessarily be assumed when comparing different assessments and that comparable assessments should not be expected to show the same level in every aspect of demand. Rather, the research should focus on investigating how different demands and different levels of demand present in different assessments balance each other out. "It is asking a lot of examiners to guarantee this balance, and a less ambitious approach requires only that the differences are made clear to everyone involved" (Pollitt *et al*., 2007, p. 166).

In this article, we give an overview of how different studies have approached the task of comparing the demands of VQs, what theories they drew from and which research methods they used to make comparisons. These studies and issues are summarised in Table 1.

Defining demand

Pollitt *et al*. (2007) define demands as "separable, but not wholly discrete, skills or skill sets that are presumed to determine the relative difficulty of examination tasks and are intentionally included in examinations/assessments" (2007, p. 196). They are inherent in the assessment tasks and are determined and built into the assessment task during the task writing process. This definition of *demands* makes them distinct from *difficulty*, which refers to how well students perform on an assessment task. While an examination question, for example, may be intended to place little demand on students, and appears to be so to the experts, in reality students may perform poorly due to some question feature overlooked by the question setter. Difficulty can be measured using performance evidence and statistics; demands can be measured only using expert judgement.

Pollitt *et al*.'s definition of demands refers primarily to the assessment task. But many studies, including awarding body studies, have taken a broader definition of demands. In many awarding body studies comparing the demands of examination question papers, mark schemes and syllabus content was a prerequisite to the comparison of performance standards. Examples can be found in the appendix of Newton *et al*. (2007). However, a purely descriptive approach, aiming only to describe various demands, "teachers – even students – might use it when choosing which qualifications to enter for, and employers [...] might use it to understand what to expect of those who have taken the exam" (Pollitt *et al*., 2007, p. 167). This type of study is particularly appropriate in situations involving new qualifications. A further step might be to attempt to quantify the relative demand of qualifications using a suitable research instrument(s). In the next sections we consider some of the methods used to describe and compare the demands of VQs.

6. Vocational Certificate of Education (VCE). This qualification had a similar modular structure to A levels and was principally taken by students of the same age. However, the qualifications were vocational. VCEs are no longer available.

7. General National Vocational Qualifications (GNVQs) were intended to offer a general introduction to an area of work. They were phased out between 2005 and 2007 (Directgov, 2009).

Table 1: Summary of studies that compare demands and include vocational qualifications

Study	*Qualifications compared*	*Theoretical framework*	*Type of demands compared*	*Focus of study*	*Research instrument*
Barry (1997)	GNVQ Science and A level Chemistry	Marton and Säljös (1976) deep versus surface learning	Teaching and learning styles, content, assessment methods	Curriculum	Participant observation, questionnaires, a test, analysis of relevant documentation
McEwen *et al*. (2001)	GNVQ and A level Science, GNVQ and A level Business	Cognitive development and expertise (Anderson, 1983, Ericcson and Smith, 1991)	Cognitive outcomes	Curriculum	Research (self-observation) diaries
Coles and Matthews (1995)	Various GNVQ and A levels in Biology, Chemistry and Physics	Bloom *et al*. (1956), Gagné (1985), Mitchel and Bartram (1994)	Subject content, general skills, type of performance or learning achievement required by stakeholders, strategies	Summative assessment	Experts identifying the qualification components and skills essential or important for their area of work
SCAA (1995)	Business Studies A level and Advanced GNVQ in Business	No theoretical framework is explicitly provided	Syllabus content, question papers, mark schemes, internal assessment tasks, teaching type and time	Summative assessment	Experts using rating scales on demand categories specified by researchers, interviews
QCA (2006a)	Personal Licence Holder Certificate across different awarding bodies	No theoretical framework is explicitly provided	Cognitive demands, test formats, test content, guided learning hours	Summative assessment	Experts looking for evidence of demand categories specified by researchers
Johnson and Hayward (2008)	Advanced Diplomas (Principal Learning component), BTEC Nationals and A levels in four different contexts: Engineering; IT; Society, Health and Development; Creative and Media	No theoretical framework is explicitly provided	Guided learning hours, content coverage, assessment models, examination requirements	Summative assessment	Experts looking for evidence of demand categories specified by researchers
Arlett (2002, 2003)*	VCE Health and Social Care across different awarding bodies	Personal construct psychology (Kelly, 1955)	Examination question papers, mark schemes, syllabus content, candidates' work	Summative assessment	KRG with rating scales. Rating scales specified by examiners
Guthrie (2003)*	A level Business Studies and VCE Business across different	Personal construct psychology (Kelly, 1955)	Examination question papers, mark schemes, syllabus content, candidates' work	Summative assessment	KRG, rating scales. Rating scales specified by examiners awarding bodies
Crisp and Novaković (2009a, 2009b)	Level 2 Certificate in Administration across centres and over time	Bloom *et al*. (1956), Hughes *et al*. (1998), Kelly's (1995) personal construct psychology	Internally assessed tasks	Summative assessment	CRAS scale, KRG, Thurstone pairs method

Notes *These studies refer to syllabus requirements, which Pollitt *et al*. (2007) refer to as demands, and therefore these studies were included.

Vocational qualification comparability research

The studies comparing VQs included in this review can be divided into two groups.

The first group comprises two studies that have taken a wide view of demands, addressing classroom practices, student learning styles and student cognition in addition to the assessment demands. In this paper, we refer to these studies as focussing on curriculum demands. There are various definitions of 'curriculum', see Nunan (1988) for a detailed discussion. We use 'curriculum' to refer to what is taught, learnt and formatively assessed, the teaching and learning experience, teaching methods, as well as the associated organisation, at the classroom, school and national level. The two studies have drawn on different theories of learning styles and student cognition.

The second group comprise the studies that have focussed primarily on summative assessment demands, such as the demands of examinations, examination questions and tasks, as well as the associated syllabus content. Some studies state that they use Bloom's taxonomy of educational objectives (Bloom *et al.*, 1956) and so a short overview of Bloom's taxonomy is provided.

Studies focussing on curriculum demands

Barry (1997) analysed the relative demands of the advanced GNVQ Science and A level Chemistry by comparing the teaching and learning approaches, content and assessment methods associated with each course, using participant observation and questionnaires. GNVQ Science was found to be more conducive to a deep approach to learning than the A level Chemistry course. Furthermore, even though the GNVQ multiple choice questions were considered easier than A level multiple choice questions, in the GNVQ test students had to achieve 70% of marks to pass whilst in the A level test only 40% was required to pass and 70% would constitute a grade A.

McEwen *et al*. (2001) compared A level with GNVQ (in Science and in Business) on three levels: pedagogy, cognitive outcomes and students' metacognition. The authors compared the classroom-based study using self-observation schedules on pedagogy and cognitive outcomes. The authors found a wide overlap in types of learning in the A level and GNVQ classrooms, with some differences. For example, in both A level and GNVQ Science classrooms, there was emphasis on applying theory to practice, problem-solving and developing skills. However, the A level put a lot of focus on memorising, understanding and consolidation, while producing new ideas and being critical were more characteristic of the GNVQ

classes. In Business A levels, memorising and consolidation were also reported but more emphasis was placed on student-centred learning than in A level Science. In the Business GNVQ the emphasis was on problem solving and decision making, as well as applying theory to practice.

Studies focussing on summative assessment demands

In the studies focussing on summative assessment demands the choice of specific demand categories was either decided by researchers in advance, or the demand categories were elicited from qualification experts. For the former, researchers drew from an established taxonomy of educational objectives and/or theories of educational and cognitive development, and/or the qualifications under investigation, and/or their experience. Coles and Matthews (1995) is an example of a study that based demand categories or themes on established literature. The method of eliciting demand types on which to compare qualifications from qualification experts was used in three awarding body comparability studies involving VQs (Arlett, 2002, 2003; Guthrie, 2003), and many studies about GQs, the most comprehensive collection of these studies is on the compact disc accompanying Newton *et al*. (2007).

In order to make comparisons Arlett and Guthrie used an initial phase inspired by Kelly's repertory grid (KRG) technique (Kelly, 1955) followed by a comparison of performance standards. The first step involves experts comparing the examination question papers, mark schemes and syllabus content from pairs of qualifications and writing down similarities and differences in demands. These ideas are then discussed and a list of construct statements together with scales for each of these statements is agreed. It is intended that the statements are about demand, and one end of each scale is the least demanding and the other end of the scale is the most demanding. A larger group of expert judges are then asked to rate qualifications on a scale for each of these construct statements. Ratings are usually from 1 to 5 or 1 to 7. Mean ratings can then be compared between syllabuses.

This is a sample of the construct statements from Arlett (2002, p. 3):

> *"Is the question paper layout accessible for candidates?"*
>
> "*To what extent are the questions readable?"*
>
> *"Questions can ask candidates to* recall *information or to* apply *knowledge. What is the relative balance of each in the question papers?"*
>
> *"Is the question structure simple or complex?"*

The studies included a question which asked about the overall level of demand of the syllabus, question papers and mark schemes.

Pollitt *et al*. (2007) argue that one issue with the KRG method is that it generates a wide range of construct statements, some of which do not refer to demands, for example, some are more descriptive, and others refer to how easy it is for the examiner to use the mark scheme. Pollitt *et al*. (2007) suggest that researchers could remove construct statements which do not refer to demands. They argue that the interviews should ask experts to describe *similarities and differences between syllabuses*, and that the interviewers should not steer the interviews to focussing on demand. However, Jankowicz (2004) holds that the interview topic can be determined by the interviewer. Therefore experts could be asked to describe *similarities and differences in demand between syllabuses*, and this might reduce the number of construct statements which are unrelated to demand.

Another method problem highlighted by Pollitt *et al*. (2007) refers to the use of scales, as different judges may apply different values or meanings to the options within the scale. For example, it is quite reasonable to question whether the mid point on a scale represents the same level of demand for a GCE examiner or a GNVQ verifier/moderator, or whether they are basing it on the level of demand of the syllabus with which they are most familiar. Pollitt *et al*. (2007) suggest using a scale from most to least demanding on which the experts rank the syllabuses. Pollitt *et al*.'s suggestion fits with KRG technique as follows. KRG involves two phases, eliciting constructs and then rating or ranking objects on the constructs (Jankowicz, 2004), in our context the objects are syllabuses. An example of a KRG study using ratings is Young *et al*. (2005) and one using rankings is Fransella and Crisp (1979).

Given these method problems, in the following section we consider studies which take a different approach, that is, experts were asked to compare qualifications using a list of demands specified in advance.

Bloom's taxonomy

Bloom's taxonomy (Bloom *et al*, 1956) classifies educational objectives within three domains: cognitive, affective and psychomotor. The taxonomy categories are ordered hierarchically, and are intended to be applicable to all types of education. The taxonomy was designed with several purposes in mind: analysing and developing standards, curricula, teaching and assessment, as well as emphasising alignment between these. It is beyond the scope of this article to discuss alignment, for further information see Maolldomhnaigh and Bealáin (1988), Prophet and Vlaardingerbroek (2003) and Liu and Fulmer (2008).

The cognitive taxonomy is divided into six categories (classes): knowledge, comprehension, application, analysis, synthesis and evaluation. Knowledge (recall of information such as facts or concepts) is the simplest and evaluation (justifying stances by judging the value of information based on a set of criteria) is the most complex. It is beyond the scope of this article to cover the behaviour categories for the affective and psychomotor taxonomies, see Krathwohl *et al*. (1964), Harrow (1972) and Simpson (1972) for details.

In the SCAA[8] (1995) report, subject experts compared the Business Studies GCE and the Advanced GNVQ in Business. They compared the syllabus content, examination question papers, mark schemes and internal assessment tasks on 1) depth and breadth, and 2) skills – *factual recall, planning, investigation, analysis and evaluation, transferability* and *application*, and rated each on a high-medium-low scale. While experts used a rating scale to compare the qualifications, it does not seem they were given examples or guidance as to what would constitute a high or low level of, for example, transferability or recall, highlighting again the problem of using rating scales as a research instrument.

QCA[9] (2006a) reports a study that compared between awarding bodies for the Personal License Holder Certificate[10] by looking into assessment practices across college, employer and training provider centres, as well as the assessment tasks. The study was detailed, covering

8. SCAA was the School Curriculum and Assessment Authority in England. It was a predecessor of the Qualifications and Curriculum Authority and the Qualifications and Curriculum Development Agency.
9. Qualifications and Curriculum Authority. The responsibilities of QCA included regulating school examinations in England.
10. Personal License Holder qualifications are intended for people who will be authorising the supply of alcohol under a Premises licence (QCA, 2006a)

the structure and format of multiple-choice tests, the assessment criteria, mark scheme, demands on candidates and other issues related to the delivery of assessments – maintenance of question item banks, mechanism for issue of results, mechanism for secure delivery, etc. It also made a clear distinction between cognitive demands of the assessment tasks and other types of test demands (text highlighting, option plausibility, reading difficulty, length of options, etc.). QCA (2006b) was a similar study about Door Supervision[11] qualifications.

Regarding the cognitive demands QCA (2006a, 2006b) used a five-level scale, with the levels being: *simple fact recall; complex recall; show understanding of a meaning: simple options; show understanding of a meaning: complex options; and apply reasoning with knowledge* (with *simple fact recall* being the lowest, and *apply reasoning with knowledge* being the highest). In these studies, the experts were not asked to rate the tests on each demand category, but simply state whether there was evidence of any of these in the assessment tasks. If experts found evidence of *simple recall*, that would constitute a demand rating of one, whereas *apply reasoning with knowledge* would constitute a rating of five.

The SCAA and the QCA studies share several features. First, they do not explicitly draw from an established theory or comparability tool. Rather they appear to use a research tool devised by the researchers from their experience. The studies do not provide an indication of the robustness of their research instrument. Secondly, the studies focus primarily on the cognitive domain, whereas the affective and psychomotor domains do not appear to be addressed. Bloom's aim was for educators to focus on all three domains, creating a more holistic form of education. Additionally, many examinations target mostly cognitive outcomes, therefore omitting some important factors, and perhaps distorting educational practice (Martinez, 1999). However, the assessment objectives of some VQs suggest that students should be able to participate in teamwork activities, develop effective communication skills, or effectively perform tasks that involve coordination or physical manipulation of tools. In this sense, any research into the demands of assessment tasks in VQs should take into account the cognitive, affective, interpersonal and/or psychomotor demands, and this has been addressed to some extent by Coles and Matthews (1995, 1998) and Johnson and Hayward (2008).

Coles and Matthews (1995) undertook a comparison of Science GQs and VQs by measuring them against the needs of HE institutions and potential employers. They used a Bloomian model as the starting point, but they adapted it using work by Gagné (1985) and Mitchel and Bartram (1994) to include the skills component, which they termed *practical capability*. The purpose of this was to recognise vocational or applied achievement. The framework they used was thus based around *recall*, *practical capability*, *interpretation*, *application*, *analysis* and *synthesis*. Coles and Matthew's (1995) work was comprehensive[12].

Johnson and Hayward (2008) compared Advanced Diplomas (Principal Learning), BTEC Nationals and A levels. The subject experts rated the requirements for several subjects including Geography, Engineering and Sociology on various issues such as: *knowledge and understanding*, *application and analysis of ideas*, *synthesis and evaluation*, *logical and critical thinking*, *literacy and language skills*, *numeracy skills*, *personal and social skills*, *learning skills*, *vocational and practical skills*. This list appears to focus on the cognitive domain. The purpose of the study was to contribute to the decision of the number of UCAS points each qualification (or each grade that can be awarded for each qualification) was assigned. UCAS points are used in university entrance procedures. Arguably universities are interested in students' cognitive skills which would explain the focus on the cognitive domain. The list above also includes personal and social skills, as well as vocational and practical skills. In this study the experts were required to note the number of times they were able to find evidence of these in the grade descriptors.

Analytic scales of demands

Bloom's taxonomy has partly influenced the development of analytic scales of demands. One such scale (Edwards and Dall'Alba, 1981), was developed in an attempt to quantify the demands placed on secondary school Science students in Australia by lessons, materials and assessments. While drawing on work by Bloom and others, the resulting scale is not a taxonomy. It identifies four categories or dimensions of demand: *complexity*, *openness*, *implicitness and level of abstraction*, and within each of these categories six levels of demand are identified. So, for example, within the complexity dimension the levels progress from *simple* operations (the lowest) to the *evaluation* as the highest. In other words, the entire Bloom's cognitive domain taxonomy is subsumed under only one dimension. The scale was designed to quantify the demands of various subjects. However, a literature search did not reveal any studies using Edwards and Dall'Alba's (1981) scale to compare VQs.

Hughes *et al*. (1998) use Edwards and Dall'Alba's scale as a starting point in developing the CRAS scale of demands. The acronym CRAS refers to the five types of demands contained within the scale:

1) *complexity* (relating to the number of components involved in a task and the relationship between these components);
2) *resources* (relating to the need to use information, either information provided or the student's own internal resources);
3) *abstractness* (the extent to which abstract ideas rather than concrete objects must be used);
4) *task strategy* (the extent to which a strategy for conducting the task must be devised by the student); and
5) *response strategy* (the extent to which a strategy for organising a response must be devised by the student).

The scale contains statements which describe the levels within each dimension, and these can be re-worded for use in different academic subjects. CRAS was developed for summarising the demands of individual assessment tasks. Greatorex and Rushton (2010) compared the CRAS scale with the frames of reference used to compare vocational demands by SCAA (1995), Coles and Matthews (1995, 1998), Arlett (2002, 2003), Guthrie (2003) and QCA (2006a, 2006b). Greatorex and Rushton (2010) concluded that CRAS was too narrow for comparing vocational syllabus demands, because it did not include some of the demands incorporated in the other studies. For instance, Coles and Matthew's (1995) include the demand "more general capabilities such as the ability to work in a team" which is primarily affective and interpersonal, whereas CRAS is predominantly concerned with cognitive demands.

11. Door supervisors are part of the security teams at public events, public houses etc. Their role includes keeping people safe and checking that only appropriate people enter the venue (Direct.gov.uk, 2010).

12. Coles and Matthews (1995) compared qualifications in a number of ways, including learning strategies. Arguably, learning strategies are a curriculum rather than a summative assessment issue and therefore this study could be classified as being in the studies comparing qualifications in terms of curriculum demands. However, we classified it as summative assessment demands as this was the focus of their research.

Crisp and Novaković (2009a, 2009b) adapted the CRAS scale for use with vocational assessment tasks by using views from VQ experts generated in a construct elicitation exercise inspired by KRG. Subsequently, the adapted CRAS scale was used to compare the demand of centre-assessed tasks across centres and over time. In order to avoid the previously mentioned problems with using rating scales, the judges were asked not to rate the assessment tasks on each dimension but to make paired comparisons of assessment tasks in terms of each dimension or type of demand. Therefore, for the purposes of this study, the scale was revised to become a framework indicating what makes tasks more or less demanding on each dimension and without a numerical scale, thereby overcoming some of the problems of rating scales.

Crisp and Novaković (2009a, 2009b) also included interviews with some of the centre tutors and students. The results of this strand of enquiry identified some differences between centres that could potentially affect assessment demands. For example, there was some indication that team tasks at one centre placed slightly greater demands on students in terms of organising their group tasks. It was also thought that working with unfamiliar peers rather than friends might alter the demands of team tasks, pointing to the affective task demands. Perhaps the most pertinent difference between tasks related to their degree of authenticity. On one hand the demands of dealing with real events, people or procedures may be higher because the task is more complex and requires more reactive skills. On the other hand, some students may find simulated tasks more demanding in terms of engaging fully with the simulated situation.

The findings of Crisp and Novaković (2009a, 2009b) highlight the need for establishing a framework of demands for VQs that would not focus exclusively on cognitive demands, but would include other types of demands such as interpersonal skills to communicate, interact and influence others to achieve goals with and through others. Studies by Coles and Matthews (1995) and by Johnson and Hayward (2008) acknowledge the limitations of taxonomies focussing only on cognitive demands by adding practical and vocational dimensions to their investigation of comparability. The 'world of work' literature also suggests that extending comparability research beyond the cognitive domain is the right course to follow. For example, McDaniel and Nguyen (2001) report that certain affective factors, such as emotional stability, agreeableness or conscientiousness correlate reasonably well with performance on certain job simulations[13]. Translated to VQs, it is easy to see how learners who have good affective skills may perform well on complex tasks adhering to occupational ethics.

Conclusion

Our review indicates that good practice for studies comparing the syllabus demand of VQs can be summarised as follows:

In the first stage, researchers would conduct KRG interviews with subject experts to elicit demands and statements of what is more and less demanding. This is similar to how many comparability studies have been conducted previously. The aim of this phase is to create a comprehensive framework which will include the cognitive domain as well as the affective, interpersonal and psychomotor domains. To facilitate the inclusion of all domains, at least a section of each interview or some interviews could be devoted to generating demands in each domain. Focussing on each domain was not a feature of many previous comparability studies.

Next, researchers would analyse the constructs into a framework of demands indicating what is more and less demanding. Pollitt *et al*. (2007) suggest that during this process researchers might need to remove some constructs which are not strictly demands.

Following the constitution of a framework several subject experts would rank two or three syllabuses from the most to the least demanding syllabus for each type of demand, thus avoiding the problems of rating scales mentioned previously. Ranking rather than rating is in line with KRG technique, and is suggested by Pollitt *et al*. (2007), but it is a departure from the common use of rating scales. Preferably, the subject experts should rank no more than three syllabuses at a time, otherwise the mental comparisons might become very challenging. The rankings can be used to calculate relative measure of demand for each type of demand.

Given that society's requirement for knowledge and skills often changes, and in turn syllabuses are reworked to reflect these changes, it is likely that any framework of demand would need to be periodically updated.

Final remarks

It was mentioned at the outset that this article was written at the beginning of a wider research project, and the wider project is still work in progress. Since the article was written the research team's thinking has shifted in two ways. First, the present article suggests using only subject experts' views about demands as the basis for a framework of demands for comparing vocational syllabus demand. But current thinking is that both subject experts' views about demands and established research literature should be used to form a framework of demands for comparing syllabus demands of units from different types of qualifications. Second, the research team's thoughts in this article were that subject experts should rank up to three syllabuses rather than use rating scales. The research team's current view is that subject experts should decide which of two units is the most demanding, for several pairs of units. In the wider research literature this process of comparing two items (of any kind) in terms of a particular characteristic is known as the method of paired comparisons. It is a research technique with a long history of use in a variety of contexts including:

- Determining the preferences of preschool children for a series of pictures of play materials (Vance and McCall, 1934).
- Weighting the seriousness of perceived health problems (McKenna *et al*. 1981).
- Comparing the demand of vocational assessment tasks (Crisp and Novaković; 2009a, 2009b).

Despite the changes in the research team's thinking, the present article usefully synthesises literature and makes several timely points.

13. Situational judgement tests are designed to measure judgement in work settings, and are intended to predict job performance. The tests present test takers with a situation(s) and a list of possible responses. The tests are a form of job simulation. See McDaniel and Nguyen (2001) for further details.

Acknowledgements

Thanks to Hannah Shiell who provided administrative support which facilitated the completion of this article.

References

Anderson, J.R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press. Cited in A. McEwen, C. McGuinness and D. Knipe. (2001). Teaching and Cognitive Outcomes in A levels and Advanced GNVQs: case studies from science and business classrooms. *Research Papers in Education* **16**, 2, 199–222.

Arlett, S. (2002). *A comparability study in VCE Health and Social Care, Units 1, 2 and 5: A review of the examination requirements and a report on the cross-moderation exercise*. A study based on the Summer 2001 examination and organised by AQA on behalf of the Joint Council for General Qualifications.

Arlett, S. (2003). A comparability study in VCE Health and Social Care, Units 3, 4 and 6: A review of the examination requirements and a report on the cross-moderation exercise. A study based on the Summer 2002 examination and organised by AQA on behalf of the Joint Council for General Qualifications.

Barry, K. (1997). An analysis of the relative demands of advanced GNVQ Science and A level chemistry. *Journal of Further and Higher Education*, **21**, 1, 43–53.

Bloom, B.S., Engelhart, M.D., Furst, E. D., Hill, W. H. & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals by a committee of college and university examiners*. New York: Longmans.

Coles, M. & Matthews, A. (1995). *Fitness for purpose. A means of comparing qualifications*. London: A report to Sir Ron Dearing.

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

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

Crisp, V. & Novaković, N. (2009b). Is this year's exam as demanding as last year's? Using a pilot method to evaluate the consistency of examination demands over time. *Evaluation and Research in Education*, **22**, 1, 3–15.

Directgov. (2009). Education and Learning. [online] Available at: http://www.direct.gov.uk/en/EducationAndLearning/QualificationsExplained/DG_10039029 (Accessed 2nd November 2009).

Directgov. (2010). Careers Advice. [online] Available at: http://careersadvice.direct.gov.uk/helpwithyourcareer/jobprofiles/JobProfile?jobprofileid=1242&jobprofilename=Door%20Supervisor&code=-1872239177 (Accessed 30th August 2010).

Edwards, J. & Dall'Alba, G. (1981). Development of a scale of cognitive demand for analysis of printed secondary science materials. *Research in Science Education*, **11**, 158–170.

Ericsson, K.A. & Smith, J. (1991). *Toward a General Theory of Expertise*. Cambridge: Cambridge, University Press. (cited in McEwen *et al.*, 2001).

Fransella, F. & Crisp, A. H. (1979). Comparisons of weight concepts in groups of neurotic, normal and anorexic females, *The British Journal of Psychiatry*, **134**, 79–81.

Gagné, R. M. (1985). The conditions of learning and theory of instruction (4th ed.). New York: Holt, Reinhart and Winston. Cited in M. Coles and A. Matthews (Eds.) 1998. *Comparing qualifications – Fitness for purpose. Methodology paper*. London: Qualifications and Curriculum Authority.

Greatorex, J. (2001). *Can vocational A levels be meaningfully compared with other qualifications?* A paper presented at the British Educational Research Association Conference, 13–15 September in Leeds, UK.

Greatorex, J & Rushton, N. (2010). Is CRAS a suitable tool for comparing specification demands from vocational qualifications? *Research Matters: A Cambridge Assessment Publication*, **10**, 40–44.

Guthrie, K. (2003). *A comparability study in GCE Business Studies, Units 4, 5 and 6 VCE Business, Units 4, 5 and: A review of the examination requirements and a report on the cross-moderation exercise*. A study based on the Summer 2002 examination and organised by Edexcel on behalf of the Joint Council for General Qualifications.

Harrow, A. (1972). *A taxonomy of the psychomotor domain: a guide for developing behavioural objectives*. New York: David McKay.

Hughes, S., Pollitt, A. & Ahmed, A. (1998). *The development of a tool for gauging the demands of GCSE and A level exam questions*. A paper presented at the British Educational Research Association conference, September in Belfast, UK.

Jankowicz, D. (2004). *The Easy Guide to Repertory Grids*. Chichester: John Wiley and Sons.

Johnson, J. & Hayward, G. (2008). *Expert group report for award seeking admission to the UCAS tariff: Advanced Diploma*. UCAS. [on line] Available at: http://www.ucas.com/documents/tariff/tariff_reports/adipreport.pdf (Accessed 17th February 2010).

Johnson, M. (2008). Assessing at the borderline: Judging a vocationally related portfolio holistically. *Issues in Educational Research*, **18**, 1. http://www.iier.org.au/iier18/johnson.html (Accessed 30th August 2010)

Kelly, G. A. (1955). *The Psychology of Personal Constructs, vols. I and II*. New York: Norton.

Krathwohl, D.R., Bloom, B. S. & Masia, B. B. (1964). *Taxonomy of Educational Objectives, Handbook II: Affective domain*. New York: David McKay.

Liu, X. & Fulmer, G. (2008). Alignment between the science curriculum and assessment in selected NY state Regents exams. *Journal of Science Education and Technology*, **17**, 374–383.

Marton, F. & Säljö, R. (1976). On qualitative difference in learning I: Outcome and process. *British Journal of Educational Psychology*, **46**, 4–11.

Maoldomhnaigh, M. Ó. & Bealáin, S. T. Ó. (1988). A comparison of the cognitive demands made by the Integrated Science Curriculum Innovation Project with those made by its written examination for the Intermediate Certificate of Education. *Irish Educational Studies* **7**, 1, 124–133.

Martinez, M.E. (1999). Cognition and the question of test item format. *Educational Psychologist*, **34**, 4, 207–218.

McDaniel, M. & Nguyen. N. (2001). Situational judgment tests: a review of practice and constructs assessed. *International Journal of Selection and Assessment*, **1**, 19–29.

McEwen, A., C. McGuinness & D. Knipe. (2001). Teaching and Cognitive Outcomes in A levels and Advanced GNVQs: case studies from science and business classrooms. *Research Papers in Education*, **16**, 2, 199–222.

McKenna, S., P., Hunt, S, M. & McEwen, J. (1981). Weighting the Seriousness of Perceived Health Problems Using Thurstone's Method of Paired Comparisons. *International Journal of Epidemiology*, **10**, 1, 93–97.

Mitchel. L. & Bartram, D. (1994). The place of knowledge and understanding in the development of the National Vocational Qualifications and Scottish Vocational Qualifications. *Competence and Assessment*, **10**, 1–47. Cited in M. Coles and A. Matthews (eds) 1998. *Comparing qualifications – Fitness for purpose. Methodology paper*. London: Qualifications and Curriculum Authority.

Newton, P., Baird, J., Goldstein, H., Patrick, H. & Tymms. P. (eds) (2007). *Techniques for monitoring comparability of examination standards*. London: QCA.

Nunan, D. (1988). Syllabus design. In: C.N. Candlin & H.G. Widdowson (Eds.), *Language Teaching: A scheme for teacher education*. Oxford: Oxford University Press.

OCR. (2009). Qualifications. [online] Available at: http://www.ocr.org.uk/qualifications/ (Accessed 2nd November 2009).

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

Prophet, R.B. & Vlaardingerbroek, B. (2003). The relevance of secondary school chemistry education in Botswana: a cognitive development perspective. *International Journal of Educational Development*, **23**: 275–289.

QCA. (2006a). *Comparability study of assessment practice: Personal license holder qualifications*, QCA/06/2709 [online] Available at: http://www.ofqual.gov.uk/files/personal_licence_holder_quals_comparability_study.pdf (Accessed 17th February 2010).

QCA. (2006b). *Comparability study of assessment practice Door supervision qualifications* QCA/06/2710 [online] Available at: http://www.ofqual.gov.uk/files/door_supervision_quals_comparability_report.pdf (Accessed 17th February 2010).

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

Simpson E. J. (1972). *The Classification of Educational Objectives in the Psychomotor Domain*. Washington DC: Gryphon House.

Vance, T. F. & McCall, L. T. (1934). Children's Preferences among Play Materials as Determined by the Method of Paired Comparisons of Pictures. *Child Development*, **5**, 3, 267–277.

Young, S.M., Edwards, H.M., McDonald, S. & Thompson, J.B. (2005). Personality Characteristics in an XP Team: A Repertory Grid Study. *SIGSOFT Software Engineering Notes*, **30**, 4, 1–7.

Rank ordering and paired comparisons – the way Cambridge Assessment is using them in operational and experimental work

Tom Bramley Research Division **and Tim Oates** Assessment Research and Development

In this article we describe the method of paired comparisons and its close relative, rank-ordering. Despite early origins, these scaling methods have been introduced into the world of assessment relatively recently, and have the potential to lead to exciting innovations in several aspects of the assessment process. Cambridge Assessment has been at the forefront of these developments and here we summarise the current 'state of play'.

In paired comparison or rank-ordering exercises, experts are asked to place two or more objects into rank order according to some attribute. The 'objects' can be examination scripts, portfolios, individual essays, recordings of oral examinations or musical performances, videos etc; or even examination questions. The attribute is usually 'perceived overall quality', but in the case of examination questions it is 'perceived difficulty'. Analysis of all the judgements creates a scale with each object represented by a number – its 'measure'. The greater the distance between two objects on the scale, the greater the probability that the one with the higher measure would be ranked above the one with the lower measure.

Background

The method of paired comparisons has a long history, originating in the field of psychophysics. Within psychology it is most closely associated with the name of Louis Thurstone, an American psychologist working in the 1920s – 1950s, who showed how the method could be used to scale non-physical, 'subjective' attributes such as 'perceived seriousness of crime', or 'perceived quality of handwriting'.

The method was introduced into examinations research in England in the 1990s principally by Alastair Pollitt, at that time Director of Research at Cambridge Assessment (then known as UCLES – the University of Cambridge Local Examinations Syndicate). He showed how the method could be used for scaling video-recorded performances on speaking tasks in the field of language testing (Pollitt and Murray, 1993), and then went on to apply it to the perennially problematic task of comparing work produced in examinations (in the same subject) from different examination boards, or from different points in time. A detailed description and evaluation of the method's use in 'inter-board comparability studies' can be found in Bramley (2007). Rank ordering is now used extensively in the comparability research work of Cambridge Assessment, and its use in operational aspects of examinations – awarding etc – is being explored and validated. But as with all approaches, it has not and will not be adopted in specific settings without testing its suitability – principally its validity and utility. This requirement for validation is in line with the standards and criteria laid down in The Cambridge Approach.

Although the mathematical details of the method can appear quite complex to non-specialists, at heart the method is very simple, the key idea being that the more times one object 'beats' another in a paired comparison, the further apart they must be on the scale. The resulting scale values are taken to be 'measures' of whatever the comparison was based on, for example 'quality of work produced'. It is assumed that, when comparing work produced in different examinations, the experts making the judgements can allow for any differences in the overall difficulty of the questions or tasks that the examinees were required to respond to.

The main theoretical attraction of the method from the point of view of comparability of examination standards is that the individual judges' personal standards 'cancel out' in the paired comparison method (Andrich, 1978). For example, a judge with a 'severe' personal standard might think that two pieces of work were both worthy of a grade B, while a judge with a more lenient personal standard might think they were both worthy of a grade A – but the two might still agree on which of the pair was better, that is, on the relative ordering of the two pieces of work.

Using the approach in research and assessment

In practice, the paired comparison method typically is very demanding – it can be extremely resource- and time-intensive. The issue for its deployment depends not least on reaching a judgement regarding its benefit-effort ratio in a specific context. In an effort to increase the efficiency of the process, Bramley (2005) showed how the same principles could be used to create a scale if the experts were asked to put several objects into a rank order rather than comparing just two. Using rankings of several objects allows many more comparisons to take place in the same time, with the advantage of allowing whole mark scales to be linked, rather than just grade boundary points. This idea of using expert judgement to link the mark scales on two (or more) tests has been the subject of a great deal of research at Cambridge Assessment, leading to several conference papers and publications (see bibliography). Black and Bramley (2008) have argued that it is a better (more valid) use of expert judgement than the method that is currently used as part of the regulator-mandated grade boundary setting process in GCSEs and A levels, and that it could have a role to play in providing one source of evidence for decisions on where to set the grade boundaries. A detailed evaluation of the rank-ordering method as a method for maintaining standards, or for investigating comparability of standards, can be found in Bramley and Gill (*in press*).

Paired comparison/rank-ordering methods have mainly been applied to the problem of comparing or maintaining standards across different tests or examinations that have been marked in the usual way. However, a far more radical use of paired comparisons/rank-ordering has been proposed by Alastair Pollitt – as an alternative to conventional marking (e.g. Pollitt, 2004; further examples in bibliography). An assumption within this is that the resulting scale is, in some situations, more valid than the raw score scale that results from conventional marking. In this scheme, both marking and standard maintaining (setting of grade boundaries) can be carried out in a single, coherent, judgement-based process. Paired comparison/rank-order judgements of work from the same examination create the scale that replaces conventional marking. Involving some pieces of work from previous examinations can 'anchor' the scale to previous scales – and hence maintain standards. In principle – although trammelled by practical problems – work from other examinations (e.g. those from other boards) could also be incorporated to ensure comparability across facets other than time.

Prototype developments in qualifications

The E-scape project led by Richard Kimbell and colleagues at Goldsmith's University (e.g. Kimbell, 2007) is a very well-funded enterprise (~ £1.8 million over its 3 stages so far) where rank-order approaches to marking are being incorporated at a larger scale than would be possible in most research exercises. The E-scape project is innovative in a number of ways, in particular for its use of technology and its attempts to achieve more valid assessment of creativity and the design process (within Design & Technology assessment). The assessment requires the creation of electronic portfolios of evidence, which are then assessed by experts using paired comparisons and rank-ordering via a customised on-line interface. So far it has been used to assess parts of GCSE Design & Technology, GCSE Geography fieldwork and GCSE science practicals (all in non-'live' pilot projects). It is also being used in several other contexts such as formative and peer assessment (see bibliography).

State of play

Innovation and openness to new ideas are fundamental to the core values of Cambridge Assessment, and the use of paired comparisons and rank-ordering in the assessment process appears to hold considerable potential. However, we are also committed to providing good evidence to support any innovations we introduce. As can be seen in the bibliography below, we have investigated and are continuing to investigate both the technical/statistical aspects of the methods, and the underlying psychology of expert judgement that they depend upon.

Research is needed in order to evaluate the quality of assessment outcomes based entirely on paired comparison or rank-order judgements, and to identify the circumstances in which these outcomes are 'better' than those produced by conventional marking. The assumptions, underlying processes, and operational issues associated with using paired comparison/rank-order judgements in public examinations require further scrutiny. Crucially, the judgement process moves more towards a 'black box' model of assessment – something which is contrary to the direction in which assessment has been developing. In addition, the increasing demand from schools, pupils and parents for detailed feedback on performance becomes problematic under such arrangements. In terms of validity, 'better' means making the case that the paired comparison/rank-order outcome supports more accurate and complete inferences about what the examinees know and can do in terms of the aims of the assessment. In terms of reliability, 'better' means showing that the paired comparison/rank-order outcomes are more replicable with different judges (markers) or different tasks (questions). In terms of practicality, we need to show that replacing marking with paired comparison/rank-order judgements is technologically, logistically and financially feasible. In terms of acceptability, 'better' means showing that examinees and other stakeholders are more satisfied with the fairness and accuracy of paired comparison/rank-order assessment outcomes, and the information from the assessment meets school, candidate and user requirements. In terms of defensibility, 'better' means showing that it is easier for examination boards, when challenged, to justify any particular examinee's result (which clearly could be a significant challenge for a system based entirely on judgement with no equivalent of a detailed 'mark scheme').

In conclusion, Cambridge Assessment is a sophisticated user and developer of rank-ordering methods and has been, and continues to be, actively involved in research into the validity of using paired comparison/rank-ordering methods in the assessment process. Our current position is that they are best deployed in standard-maintaining contexts, when the assessments being compared are as similar as possible (e.g. examinations from the same board in the same subject in consecutive examination sessions). We are actively exploring their applicability to more general investigations of comparability and to mainstream qualifications and assessments.

Bibliography

Paired comparison and rank-ordering methods

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, **2**, 449–460.

Bell, J.F., Bramley, T. & Raikes, N. (1998). Investigating A-level mathematics standards over time. *British Journal of Curriculum and Assessment*, **8**, 2, 7–11.

Black, B. (2008). *Using an adapted rank-ordering method to investigate January versus June awarding standards*. Paper presented at the Fourth Biennial EARLI/ Northumbria Assessment Conference, Berlin, Germany, August 2008.

http://www.cambridgeassessment.org.uk/ca/collection1/digitalAssets/186335_BB_Jan_June_EARLI.pdf [Accessed 24/6/10].

Black, B. & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education*, **23**, 3, 357–373.

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgement. *Journal of Applied Measurement*, **6**, 2, 202–223.

Bramley, T. (2007). Paired comparison methods. In: P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. 246–294. London: Qualifications and Curriculum Authority. [Contains many references to studies that have used the paired comparison method in inter-board comparability research]

Bramley, T. (2009). *The effect of manipulating features of examinees' scripts on their perceived quality*. Paper presented at the annual conference of the Association for Educational Assessment – Europe (AEA-Europe), Malta, November 2009.

Bramley, T. (2010). *Locating objects on a latent trait using Rasch analysis of experts' judgements*. Paper presented at the conference "Probabilistic models for measurement in education, psychology, social science and health", Copenhagen, Denmark, June 2010. http://www.cambridgeassessment.org.uk/ca/collection1/digitalAssets/186233_TB_locating_objects_Rasch2010.pdf

Bramley, T., Bell, J.F. & Pollitt, A. (1998) Assessing changes in standards over time using Thurstone Paired Comparisons. *Education Research and Perspectives*, **25**, 2, 1–23.

Bramley, T. & Black, B. (2008). *Maintaining performance standards: aligning raw score scales on different tests via a latent trait created by rank-ordering examinees' work*. Paper presented at the Third International Rasch Measurement conference, University of Western Australia, Perth, January 2008. http://www.cambridgeassessment.org.uk/ca/collection1/digitalAssets/171143_TB_BB_rank_order_Perth08.pdf [Accessed 24/6/10].

Bramley, T. & Gill, T. (*in press*). Evaluating the rank-ordering method for standard maintaining. *Research Papers in Education*.

Curcin, M., Black, B. & Bramley, T. (2009). *Standard maintaining by expert judgement on multiple-choice tests: a new use for the rank-ordering method*. Paper presented at the British Educational Research Association annual conference, University of Manchester, September 2009. http://www.cambridgeassessment.org.uk/ca/collection1/digitalAssets/184999_BERA_paper_Curcin_Black_and_Bramley.pdf [Accessed 24/6/10].

Gill,T., Bramley, T. & Black, B. (2007). *An investigation of standard maintaining in GCSE English using a rank-ordering method*. Paper presented at the British Educational Research Association annual conference, Institute of Education, London. http://www.cambridgeassessment.org.uk/ca/collection1/digitalAssets/186333_TG_Eng_rankorder_BERA_paper_final. pdf [Accessed 24/6/10].

Greatorex, J., Novaković, N. & Suto, I. (2008). *What attracts judges' attention? A comparison of three grading methods*. Paper presented at the Annual Conference of the International Association for Educational Assessment, Cambridge, September 2008. http://www.cambridgeassessment.org.uk/ca/collection1/digitalAssets/174284_What_attracts_judges__attention.pdf [Accessed 24/6/10].

Novaković, N. & Suto, I. (2009). *How should grade boundaries be determined in examinations? An exploration of the script features that influence expert judgements*. Paper presented at the European Conference for Educational Research, Vienna, Austria, September 2009.

Novaković, N., and Suto, I. (2010). The reliabilities of three potential methods of capturing expert judgement in determining grade boundaries. *Research Matters: A Cambridge Assessment Publication*, **9**, 19–24.

Pollitt, A. & Murray, N.J. (1993). What raters really pay attention to. Language Testing Research Colloquium, Cambridge. Reprinted in: M. Milanovic, & N. Saville (Eds.) *Studies in Language Testing 3: Performance Testing, Cognition and Assessment*. Cambridge University Press: Cambridge.

Raikes, N., Scorey S. & Shiell, H. (2008). *Grading examinations using expert judgements from a diverse pool of judges*. Paper presented at the Annual Conference of the International Association for Educational Assessment, Cambridge, September 2008. http://www.cambridgeassessment.org.uk/ca/collection1/digitalAssets/186337_Raikes__Scorey_and_Shiell__IAEA_2008_-Grading_examinations_using_expert_judgements_from_a_diverse_pool_of_judges.pdf [Accessed 24/6/10].

Thurstone, L.L. (1959). *The measurement of values*. Chicago: University of Chicago Press.

"Let's stop marking exams"

Pollitt, A. (2004). *Let's stop marking exams*. Paper presented at the 30th Annual Conference of the International Association for Educational Assessment, June, Philadelphia, USA. [online] Available at: http://www.cambridgeassessment.org.uk/ca/digitalAssets/113942_Let_s_Stop_Marking_Exams.pdf [Accessed 25/02/10]

Pollitt, A. (2009). *Abolishing marksism and rescuing validity*. Paper presented at the 35th Annual Conference of the International Association for Educational Assessment, September, Brisbane, Australia. [online] Available at: http://www.iaea2009.com/abstract/69.asp [Accessed 24/02/10]

Pollitt, A. & Crisp, V. (2004). *Could Comparative Judgements of Script Quality Replace Traditional Marking and Improve the Validity of Exam Questions?* Paper presented at the British Educational Research Association Annual Conference, September, UMIST, Manchester, UK. [online] Available at: http://www.cambridgeassessment.org.uk/ca/digitalAssets/113798_Could_comparative_judgements.pdf [Accessed 25/02/10]

Pollitt, A. & Elliott, G. (2003). *Finding a proper role for human judgement in the examination system*. Paper presented at the Qualifications and Curriculum Authority Seminar on 'Standards and Comparability', April 2003. [online] Available at: http://www.cambridgeassessment.org.uk/ca/digitalAssets/113954_Finding_a_Proper_Role_for_Human_Judgement_in_the_Examination.pdf [Accessed 25/02/10]

E-scape project

Davies, D. (2008). *E-scape phase 3 science final report – December 2008*. [online] Available at: http://www.bathspa.ac.uk/schools/education/research/docs/08-09/09-e-scape-final-report.pdf [Accessed 24/02/10]

Kimbell, R. (2007). E-assessment in project e-scape. *Design and Technology Education: An international journal*, **12**, 2, 66–76.

Kimbell. R., Brown Martin, G., Wharfe, W., Wheeler, T., Perry, D., Miller, S., Shepard, T., Hall, P. & Potter, J. (2005). *E-scape e-portfolio assessment. Phase 1 report*. [online] Available at: http://www.gold.ac.uk/static/teru/pdf/e-scape1.pdf [Accessed 24/02/10]

Kimbell, R., Wheeler, T., Miller, S. & Pollitt, A. (2007). *E-scape e-portfolio assessment. Phase 2 report*. [online] Available at: http://www.gold.ac.uk/media/e-scape2.pdf [Accessed 24/02/10]

Kimbell, R., Wheeler, T., Stables, K., Shepard, T., Martin, F., Davies, D., Pollitt, A. & Whitehouse, G. (2009). *E-scape e-portfolio assessment. Phase 3 report*. [online] Available at: http://www.gold.ac.uk/media/e-scape_phase3_report.pdf [Accessed 24/02/10]

Martin, F. (no date). *E-scape: Briefing paper for Secondary Phase Committee* [online] Available at: www.geography.org.uk/download/GA_PREscapeBriefingDoc.pdf [Accessed 24/02/10]

Martin, F. & Lambert, D. (2008). *Report for geography 'e-scape' Initial Trial. June 2007–December 2008*. [online] Available at: http://www.geography.org.uk/download/GA_PREscapeFinalReport.pdf [Accessed 24/02/10]

Newhouse, P., Miller, D., Campbell, A. & Cooper, M. (2009). *Digital forms of assessment. Extra report on the analysis of 2008 data. Applied Information Technology. A report for the Curriculum Council of Western Australia*. [online] Available at: http://csalt.education.ecu.edu.au/downloads/AIT_Report2008.pdf [Accessed 24/02/10]

Technology Education Research Unit (2009). *E-scape project information*. [online] Available at: http://www.gold.ac.uk/teru/projectinfo/ [Accessed 24/02/10]

Technology Education Research Unit and CIE (undated). Video of the IGCSE English Experiment in E-assessment http://www.cie.org.uk/qualifications/academic/middlesec/igcse/video/IGCSE_english_experiment_esssessment [Accessed 24/02/10]

A better approach to regulating qualification standards

Bene't Steinberg Group Director, Public Affairs **and Sarah Hyder** Parliamentary Manager, Public Affairs

In light of the forthcoming Government White Paper on education due out in Autumn 2010, Cambridge Assessment explains here how new patterns of engagement between those concerned with the creation and use of assessments can lead to the better regulation of public examinations. This viewpoint was posted on the Cambridge Assessment website in September 2010.

The question of standards

The original purpose of public examinations (created in the mid-nineteenth century, mainly by universities) was to drive up standards at the lower levels of education and provide a stream of potential undergraduates. Grammar Schools and the Headmasters' and Headmistresses' Conference (HMC) Schools used them to certificate the learning being delivered. Subsequently, the Government required them to ascertain it was getting value for the money it spent on schools. That original purpose still stands today.

Exams have become crucial both for entry to a Higher Education (HE) sector taking nearly 50% of the cohort each year and for securing the bulk of jobs with progression prospects. In the late 1990s a more businesslike attitude took root among the exam boards, a school accountability framework based on exam results was introduced and, in 2002, a commercial exam board was introduced into the system.

This led to fears that boards could be lowering standards in order to achieve market share. The reality is that the aggregate market share of the boards has remained remarkably constant since the introduction of Curriculum 2000. Nevertheless, the question for the new century has changed from measuring whether education is good via examinations to whether the examinations in themselves are a good measure of that education.

Why there is a problem

Over the past forty years, exam boards became ever more concerned with technical accuracy while 'users' of qualifications such as HE and employers became more concerned with other issues. As a consequence, the British state ended up disintermediating subject communities, HE, professional societies, employers, teachers and those developing and providing examinations by taking upon itself the role of defining the content of syllabuses and the way in which they were examined. Thus, 'users' were divorced from 'producers'. Producers have continued to carry out a difficult and arcane task with ever increasing accuracy but with little direct contact with users to help them re-balance that precision with some healthy macro overviews of the purpose of the exercise.

The current situation

The last Government sought to address the question of standards by setting up a new regulator, the Office of Qualifications and Examinations Regulation (Ofqual), which has a more clearly defined role than its predecessor, the Qualifications and Curriculum Authority (QCA). The Coalition Government has made it clear that it does not regard this as being the best way of ensuring standards are maintained and has committed to legislation giving Ofqual the powers it needs to enforce rigorous standards.

Ministers have already stated that they are not interested in the direct regulation of 'products' and are abolishing Ofqual's partner quango, the Qualifications and Curriculum Development Agency (QCDA). The QCDA is currently responsible for defining qualification (design) criteria – such as the number of units, the grading structure and methods of assessment – and subject (content) criteria.

The regulator is likely to be most effective if it is allowed to focus on a specific objective, rather than a collection of objectives which it currently holds. Narrow and deep regulation creates a more effective regulator than a broad and superficial approach.

How can standards best be maintained?

1 Users need to take the major role in specifying the content criteria of qualifications – enabling them to help set the standards.

2. Exam boards need to agree between themselves on design criteria – enabling them to set and maintain the standard in relation to each other.

3. 'Communities of practice' (see below) need to be set up around each qualification – enabling the standards of each qualification to be owned and maintained by all those with a direct interest in them.
4. The regulator must focus on standards alone rather than its other current objectives. Its role in this system would be to underpin inter-exam board agreements as well as those between boards and users.

The best international qualifications – the International Baccalaureate (IB), the Pre-U, the International General Certificate of Secondary Education (IGCSE) – are such because they have a minimum of state intervention, with producers and users of the qualification creating a community of ownership and practice that takes upon itself the responsibility for maintaining the integrity of the qualification.

If one gives users of qualifications a leading role in determining the content of those qualifications and creates communities of practice which include those users, the role that the regulator then plays can be redefined to better serve the nation's needs. Users are given a direct stake in maintaining the standard and a community is created that is bound to that standard. Therefore, the regulator goes from being a mediator between the users and producers of qualifications which makes its own decisions on the standard, to allowing those with an interest in maintaining the standard a greater role in doing so.

Not all of this requires legislation – but all parties must agree to meet their obligations as outlined below in order for a new regulatory approach to work.

The Sykes Report[1] stated "The primary determinants of the content …of A levels should be the requirements of the subject and of the users of the qualification", with which Cambridge Assessment agrees. The same approach holds good for Level 2 qualifications (GCSE and others), with the users being subject communities and learned societies, schools and colleges managing progression, and businesses where appropriate.

Subject professionals then take the major role in determining the knowledge, skills and understanding they expect of a candidate in a subject (academic or vocational) at that point in their learning. They also continue to engage actively with awarding bodies over the lifetime of the qualification. Those professionals therefore have a direct interest in preserving the currency of the specific qualification in which they are involved – and a method by which they can ensure that the currency is upheld.

For the purposes of A level, it makes sense for the Government and HE to make clear that the primary purpose of A levels is for HE entry; this sends signals to the exam-taking cohort as to which qualifications are worth taking and that HE is prepared to take a major role in preserving the currency of the qualifications used for entry to it.

Users also have a role in suggesting design criteria but assessment expertise is primarily located within exam boards. Thus, design criteria are best developed by those experts in assessment working in close consultation with the teaching community, the subject community and users. In this way, the users' preferences would be taken into account, set against the practicalities of assessment practice (time, cost, question type in relation to knowledge, and so forth).

Different subjects may well choose different styles of examination that most suit the teaching and assessment of their subject (e.g. some favouring a linear approach, others two units, three or more). Given that the users would be the guarantors of the standard, there is no need for a regulator to insist on direct similarities in interests of bureaucratic symmetry.

In order to hold the standard over time, it is vital that 'communities of practice' are created. That is, "groups of people who share…a passion for something they do and learn how to do it better as they interact regularly…Membership…implies a commitment to the domain, and therefore a shared competence that distinguishes members from other people… [they] engage in joint activities and discussions, help each other, and share information…build relationships that enable them to learn from each other. They develop a shared repertoire of resources…This takes time and sustained interaction."[2]

Qualification communities of practice bring together leading users, subject specialists, teachers, syllabus designers and question writers to share a particular view of what constitutes the standard in relation to a subject level. Because they work together, continually improving their understanding, they own the standard and protect it in on a day to day basis against the vagaries of pedagogical or political fashion. This is the way in which the IGCSE, the IB and the Cambridge Pre-U manage standards – without the agency of the state.

By giving awarding bodies greater ownership over the development of qualification and subject criteria, they become more accountable to users and to the general public. Rather than acting as a conduit between the state's requirements and the end user, which confuses the accountability process, direct interaction with users means that awarding bodies are incentivised to be more accountable to those end users – and therefore to the wider public.

With this approach, the regulator is set free to focus wholly upon standards and the protection of the public from the production of worthless qualifications. For these purposes it requires only Objective (a) as set out in the Apprenticeships, Skills, Children and Learning (ASCL) Act 2009 – the qualifications standards objective[3] – but only through maintaining the standards of the bodies that award qualifications.

The responsibility for maintaining standards of the bodies that award qualifications can be undertaken by setting criteria for systems, structures, procedures, quality assurance and continuous improvement, and licensing the organisation for the production of one or more types of qualification on the grounds that it meets the criteria. The responsibility for setting the standard of each individual qualification would be taken up by the user group and the maintenance of it undertaken by communities of practice that necessarily include those users. They would, in legal language "ensure that qualifications give a reliable indication of knowledge, skills and understanding". Therefore, Ofqual would no longer accredit individual qualifications at this level. Provided the awarding body had the engaged support of a sufficient number of users (the number set possibly through regulation) and had a community of practice, or had plans to create one, Ofqual would merely register the qualification.

Ofqual would have a role in signing off the design criteria as agreed jointly by the awarding bodies in consultation, for the sole purpose of ensuring a level of equivalency of qualifications within each subject (not between subjects). If a user group favoured a course which did not fit with these agreed qualification criteria, it would make a case for a derogation from the norm. Ofqual would also have a role in ensuring that the awarding body was holding its standard over the lifetime of the

1 http://www.conservatives.com/News/News_stories/2010/03/~/media/Files/Downloadable%20Files/Sir%20Richard%20Sykes_Review.ashx

2. Wenger, Etienne (2006). Communities of Practice: A brief introduction. http://www.vpit.ualberta.ca/cop/doc/wenger.doc

3. The qualifications standards objective – to secure that regulated qualifications give a reliable indication of knowledge, skills and understanding, and indicate a consistent level of attainment (including over time) between comparable regulated qualifications.

qualification through active management of its community of practice. It would also undertake most of the national, intra-UK and international comparability studies required to keep England's qualifications at the forefront of international practice. The regulator's responsibilities would therefore become more about monitoring the standard over time, as well as having the powers to instruct people to move back to the standard they set on the first iteration of the qualification if they have shifted from it.

Legislative outcomes

The Regulator

This approach focuses the regulator on the key issue of standards in public qualifications. There are other reasons for removing some of its other objectives.

Professor Alison Wolf, Professor of Public Sector Management at King's College, London, makes it clear[4] that the principal tools of regulation in education are:

1. Initial and permanent licensing of providers
2. Regular re-licensing of providers
3. Inspection
4. Publishing quantitative measures of individual providers' output and/or quality
5. Direct control and regulation of products and/or delivery mechanisms

Professor Wolf writes extensively on which of these tools work best. Her analysis makes it clear that at least three of Ofqual's objectives can be secured in better ways than via an exams regulator.

- The system laid out above requires Ofqual to continue to have the powers given to it under Objective (a), the qualifications standards objective[5]. This is best done by ensuring awarding bodies are 'fit and proper' providers through licensing or re-licensing procedures – not by looking on the standard of every individual qualification. It therefore does not require Sections 138–140.
- Ofqual's assessments standards objective (b)[6] relates exclusively to statutory national curriculum and Early Years Foundation Stage assessments and we would agree that Ofqual should continue to have a role in maintaining the standard of these assessments. We would argue that because qualifications communities of practice are unlikely to form to the same extent around the content and form of assessments at this level, Ofqual needs to maintain a role in this area.
- Ofqual's public confidence objective (c)[7] is best achieved by performing the task, as with all other regulators, of upholding standards. To have a specific objective like this encourages the employment of ever larger communications teams delivering ever more communications programmes rather than a commitment to proper investigation and research. The re-linking of HE, business and subject communities directly with awarding bodies means that the users of qualifications give or withhold their support and the confidence of the public in the ability of a qualification to deliver progression is assured without the need for a regulator to engage in PR activities. The regulator does not need to build its own reputation – the qualifications should build their own reputation through the recognition of their users.
- Ofqual's awareness objective (d)[8] seems to replicate and place into a central structure the marketing operations of the awarding bodies which seek to bring attention to their individual qualifications and the benefits of them. A vibrant market is the best guarantee of public awareness of the available opportunities. In addition, the past decade has seen the creation of large numbers of comparison websites which might take on this role, or be encouraged to do so. UCAS could use its knowledge to provide such a service, particularly if Higher Education and sectoral business groups rise to the challenge of the Minister for Higher Education and start to send out clearer signals as to which qualification is best for their purposes. The Objective also leads to loss of focus and requires additional resources. The previous government's determination to bring all those businesses that provide their own or sectoral qualifications into the regulated sector essentially marketed the regulator's role. We would submit that if the regulator succeeds in establishing itself as competent in its main role, such potential market entrants will find their own way to it.
- Ofqual's efficiency objective (e)[9] is best secured through proper competition. Ofqual has commissioned six market studies so far, none of which has indicated the making of extraordinary profit/surplus by any agents. The 'markets' are many and varied, while the provision of 'free' services[10] attached to qualifications is an immensely complicated arena. We would submit that the Competition Commission or the Office of Fair Trading has a far wider and deeper knowledge of complex markets than Ofqual can ever match. This area of responsibility could therefore usefully be transferred to either of them, with a consequent reduction in the cost of the regulator.

Users

The structures which will encourage users to engage will need to be laid out – either by Order or through primary legislation. For example, in the case of HE, the Quality Assurance Agency (QAA) might include engagement with awarding bodies as one of its criteria for defining a 'quality Higher Education Institution', or the QAA Code of Practice on Programme Design (Section 7) could usefully include reference to the need to take note of the incoming knowledge and skills of students when designing a course.

4. Wolf, A. (2010). *How to shift power to learners*. London: LSN Centre for Innovation in Learning. https://crm.lsnlearning.org.uk/user/order.aspx?code=100006
5. The qualifications standards objective – to secure that regulated qualifications give a reliable indication of knowledge, skills and understanding, and indicate a consistent level of attainment (including over time) between comparable regulated qualifications.
6. The assessments standards objective – to promote the development and implementation of regulated assessment arrangements which give a reliable indication of achievement, and indicate a consistent level of attainment (including over time) between comparable assessments.
7. The public confidence objective – to promote public confidence in regulated qualifications and regulated assessment arrangements.
8. The awareness objective – to promote awareness and understanding of the range of regulated qualifications available, the benefits of regulated qualifications to learners, employers and the higher education sector, and the benefits of recognition to bodies awarding or authenticating qualifications.
9. The efficiency objective – to secure that regulated qualifications are provided efficiently and in particular that any relevant sums payable to a body awarding or authenticating a qualification represent value for money.
10. e.g. training, syllabus provision, teaching tools.

Elsewhere, the Higher Education Funding Council for England (HEFCE) might inherit some of the money saved from the abolition of the QCDA for funding engagement activities, similar to the Aimhigher programme[11] – and of a similar order. It is likely that a small funding stream will need to be made available in order that universities can allow staff adequate time to engage in this process, thereby ensuring 'quality' rather than 'tick box' engagement.

It may be that seconding academics to awarding bodies during the early stages of the design process to ensure the standard was properly set would be a good use of seedcorn monies. Certainly, continuous engagement from early design through to production will require some element of incentivisation, given the vast range of other duties expected of the modern academic.

The impact criteria of the Research Assessment Exercise (RAE) could also provide a helpful lever. There is a perfectly reasonable case to be made that disseminating knowledge to the next level down of the education system is nearly as important as some other RAE criteria. Clearly, it would not rate as importantly as an academic paper but is of great importance to the long-term health of the nation.

And it may well be that the HE Academy could usefully turn its mind to how it might provide a service both to HE and wider education by providing structures and resource to encourage such engagement.

Stability

A current unhelpful part of the process is the frequency of qualification 'accreditation cycles'. The frequency of these changes is driven by regulatory pressures rather than by a change in the structure and content of knowledge in subject areas, change in effective pedagogy, evidenced innovation in curriculum practices, or emerging needs in the learner group. None of these factors work to particular timescales.

Because the reaccreditation process occurs on a frequent basis and requires the change of a qualification across all subject areas, awarding bodies are required to engage across the whole of the user group and in a limited period of time. This reduces the likelihood of quality engagement. In addition, repeated changes to qualifications which are beyond and more frequent than those necessitated by subject and pedagogical change, as mentioned above, can have a negative impact on maintaining the standard of qualifications.

Regulatory engagement ought to be based on a presumption in favour of stability which should prevail over the current approval process of synchronised accreditation to ensure compatibility across boards. A General Duty under the current Section 129 (1)[12] would embed a more acceptable approach into the process.

In summary, our thesis is that: standards of qualifications are better maintained if they are owned by the users and deliverers rather than through a bureaucratic process. If this responsibility is returned to users and communities of practice, minimal and useful regulation can then follow.

11. Aimhigher is a national programme which aims to widen participation in higher education by raising HE awareness, aspirations and attainment among young people from under-represented groups. http://www.aimhigher.ac.uk/practitioner/programme_information/about_aimhigher.cfm

12. Section 129 General duties: (1) So far as is reasonably practicable, in performing its functions Ofqual must act in a way – (a) which is compatible with its objectives, and (b) which it considers most appropriate for the purpose of meeting its objectives.

EXAMINATIONS RESEARCH

Statistical Reports

The Statistics Team Research Division

The ongoing 'Statistics Reports Series' provides statistical summaries of various aspects of the English examination system such as trends in pupil attainment, subject uptake, qualifications choice and subject provision at school. These reports, produced using national-level examination data, are available on the Cambridge Assessment website:

http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Statistical_Reports

The following reports have been published since Issue 10 of *Research Matters*:

- Statistics Report Series No. 20: How old are GCSE candidates?
- Statistics Report Series No. 21: A-Level uptake and results by gender, 2002–2009
- Statistics Report Series No. 22: GCSE uptake and results by gender, 2002–2009
- Statistics Report Series No. 23: A-Level uptake and results by school type, 2002–2009
- Statistics Report Series No. 24: GCSE uptake and results by school type, 2002–2009

Research News

Conferences and seminars

International RASCH Conference 2010

In June Tom Bramley attended the International Rasch Conference in Copenhagen, Denmark, and presented a paper on: *Locating objects on a latent trait using Rasch analysis of experts' judgments*.

International Association for Educational Assessment (IAEA)

The 36th IAEA Annual Conference took place in Bangkok, Thailand, from 22nd–27th August 2010. The theme of the conference was 'Assessment for future generations'. Colleagues from Cambridge Assessment presented the following papers:

Rebecca Hopkin, Martin Johnson, Hannah Shiell, John F. Bell and Nicholas Raikes: *Marking advanced extended essays on screen and on paper: is overall marking accuracy reliable across marking modes?*

Victoria Crisp and Stuart Shaw: *How hard can it be? Issues and challenges in the development of a validation method for traditional written examinations*.

Louis Yim: *A comparison between the effect of using pseudo-candidates' scripts and real-candidates' scripts in a rank-ordering comparability methodology at syllabus level*.

Beth Black: *Investigating seeding items used for monitoring on-line marking: factors affecting marker agreement with the gold standard marks*.

Stuart Shaw and Irenka Suto: *A tricky task for teachers: Assessing pre-university students' research reports*.

European Conference on Educational Research (ECER)

In August Irenka Suto attended the ECER conference in Helsinki, Finland. Over 2000 delegates attended the conference, which took place over three days and comprised 27 different networks. The theme was 'Education and cultural change'. Irenka presented two papers:

Irenka Suto, Beth Black and Tom Bramley: *The Interrelations of Features of Questions, Mark Schemes and Examinee Responses and their Impact upon Marker Agreement*.

Stuart Shaw and Irenka Suto: *A tricky task for teachers: Assessing pre-university students' research reports*.

EARLI/Northumbria Assessment Conference 2010

Rebecca Hopkin attended the Fifth Biennial Northumbria/EARLI SIG Assessment Conference in Northumberland in September. Within the general theme of 'Assessment for Learners', the conference programme covered the following areas:

- Formative and summative assessment to improve learning
- Assessment: consequences and contexts for learners
- Learner achievements and assessment

British Educational Research Association (BERA)

The BERA Annual Conference was held from 1st–4th September at the University of Warwick. Colleagues from the Research Division and CIE presented the following papers:

Carmen Vidal Rodeiro and Rita Nádas: *The effects of the new modular GCSE examinations on students' outcomes, motivation and workload*.

Joanne Emery, Elizabeth Sykes, Tim Oates, John F. Bell and Carmen Vidal Rodeiro: *A review of the birth date effect on educational attainment in England*.

Tim Gill: *An analysis of examination uptake and performance of schools in the academies programme*.

Beth Black and Milja Curcin: *Group dynamics in determining 'gold standard' marks for seeding items and subsequent marker agreement*.

Stuart Shaw and Martin Johnson: *Towards an understanding of the impact of annotations on returned examination scripts*.

Hannah Shiell and Irenka Suto: *Influences on moderation and standards maintenance in school-based summative assessment: how do professional concerns differ from the evidence?* (Poster)

IQB IV European Congress of Methodology

In July John Bell attended the IQB IV European Congress of Methodology in Potsdam, Germany and presented a poster: *The empty file drawer: An explanation of the small study effect*.

Association for Educational Assessment (AEA) – Europe

The theme for the 11th AEA-Europe Conference, which took place in Oslo, Norway in November, was 'Managing assessment processes: policies and research'. Colleagues from Cambridge Assessment presented the following papers:

Tim Oates: *If at first you don't succeed, try, try again. Using reform of qualifications to effect structural change in vocational training arrangements*.

Sylvia Green and Victoria Crisp: *A new model of assessment for 14 to 19 year olds: What do students and their teachers think of Diploma assessments?*

Milja Curcin, Beth Black and Tom Bramley: *Towards a suitable method for standard-maintaining in multiple-choice tests: capturing expert judgement of test difficulty through rank-ordering*.

John F. Bell: *A comparison between modular and linear examinations in secondary education: the impact of maturational effects and regular feedback on performance and motivation*.

Martin Johnson: *Marking advanced extended essays on screen and on paper: Is overall marking accuracy reliable across marking modes?*

Stuart Shaw: *Issues around how best to provide evidence for assessment validity*.

Stuart Shaw and Victoria Crisp: *How valid are A levels? Findings from a multi-method validation study of an international A level in geography*.

Louis Yim and Mark Dowling: *A benchmarking exercise between examination boards using a rank-ordering methodology at syllabus level*.

Stuart Shaw and Victoria Crisp: *Identifying a set of methods for validating traditional examinations: A difficult task requiring multiple methods*. (Poster)

Publications

Sylvia Green was invited to be guest editor of a Special Issue of *Research Papers in Education* on contemporary issues in assessment. Irenka Suto was the deputy guest editor. *Research Papers in Education: Policy and Practice*, **25**, 3 was published in September.

The following articles have been published since Issue 10 of *Research Matters*:

Bramley, T. & Gill, T. (2010). Evaluating the rank-ordering method for standard maintaining. *Research Papers in Education: Policy and Practice*, **25**, 3, 293–317.

Johnson, M. and Burdett, N. (2010). Intention, interpretation and implementation: some paradoxes of Assessment for Learning across educational contexts. *Research in Comparative and International Education*, **5**, 2, 122–131.

Johnson, M. and Burdett, N. (2010). School-based assessment in international practice. *Problems of Modern Education*, **4** (Russian journal) http://www.pmedu.ru/res/2010_4_6.pdf

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

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

ISSN: 1755–6031